P4 Project by William Thomas

The purpose of this project is to use exploratory data analysis techniques which is required by Udacity's Data Analyst Nanodegree requirements of the red wine dataset. Before this project, I had no previous knowledge of red wine and the specific criteria that recognizes it as good or bad quality. The dataset can be downloaded here(https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub?embedded=true (https://docs.google.com/document/d/1qEcwltBMlRYZT-l699-71TzInWfk4W9q5rTCSvDVMpc/pub?embedded=true)). This project will include the red wine dataset for analysis by exploring the relationships between different variables.

```
## [1] "/Users/williamthomas/Downloads"
```

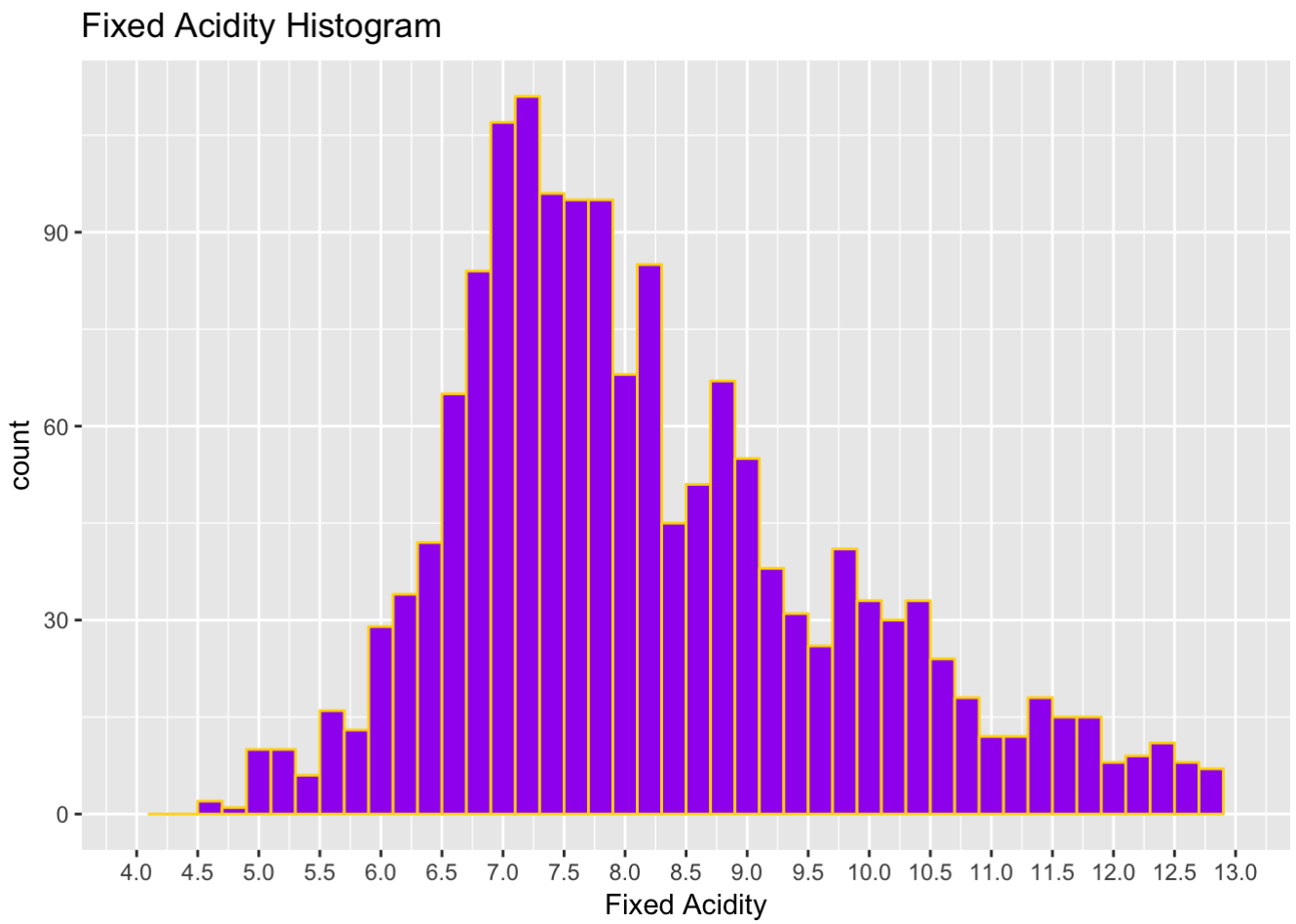
Univariate Plots Section

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
##        X          fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol         quality
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

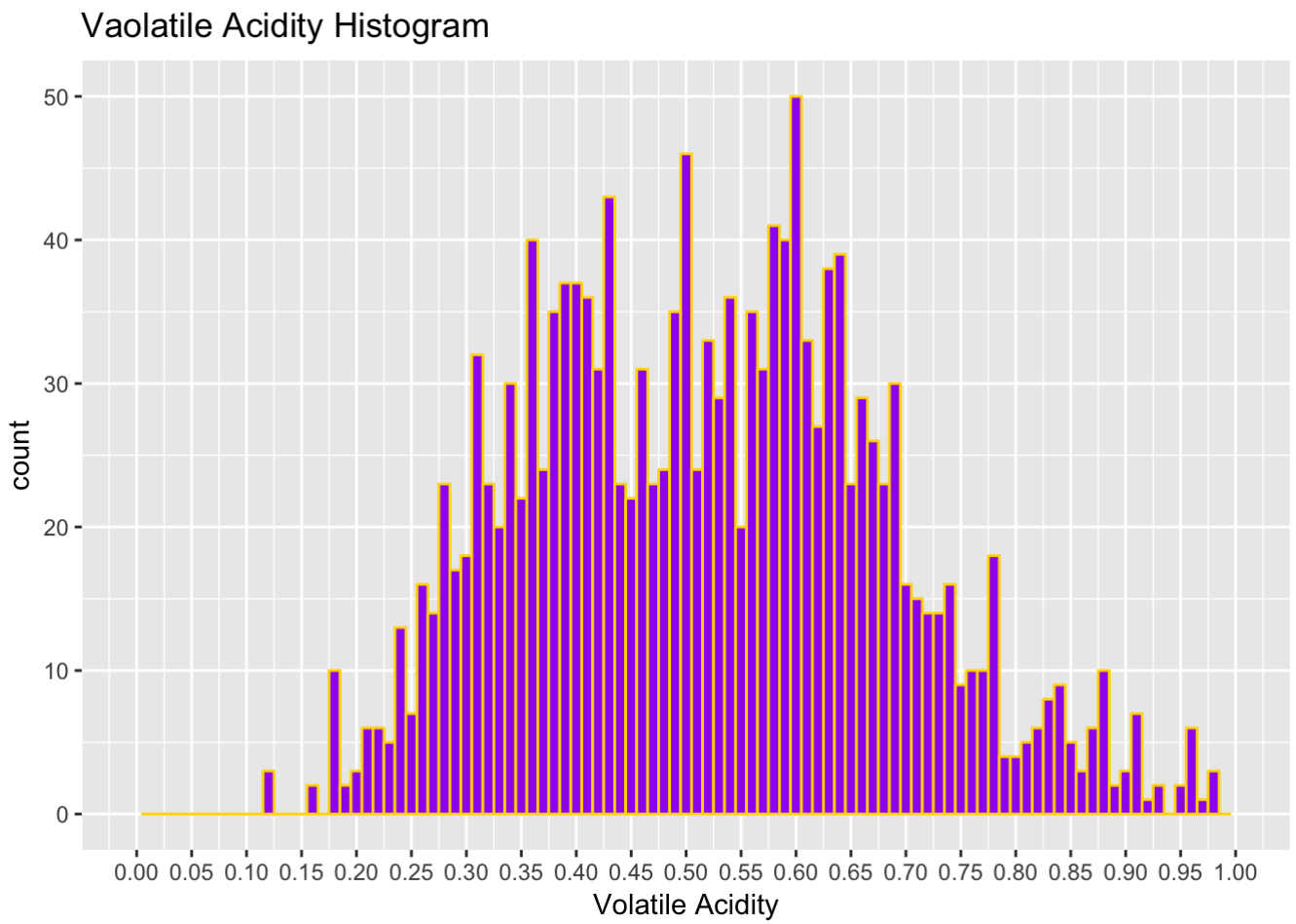
```
## Warning: Removed 20 rows containing non-finite values (stat_bin).
```



Fixed Acidity has a mean of 8.32 and a median of 7.90. The histogram seems to be somewhat distributed to the right. The max fixed Acidity was 15.90, an outlier, which was taken out of the data set when plotting histogram.

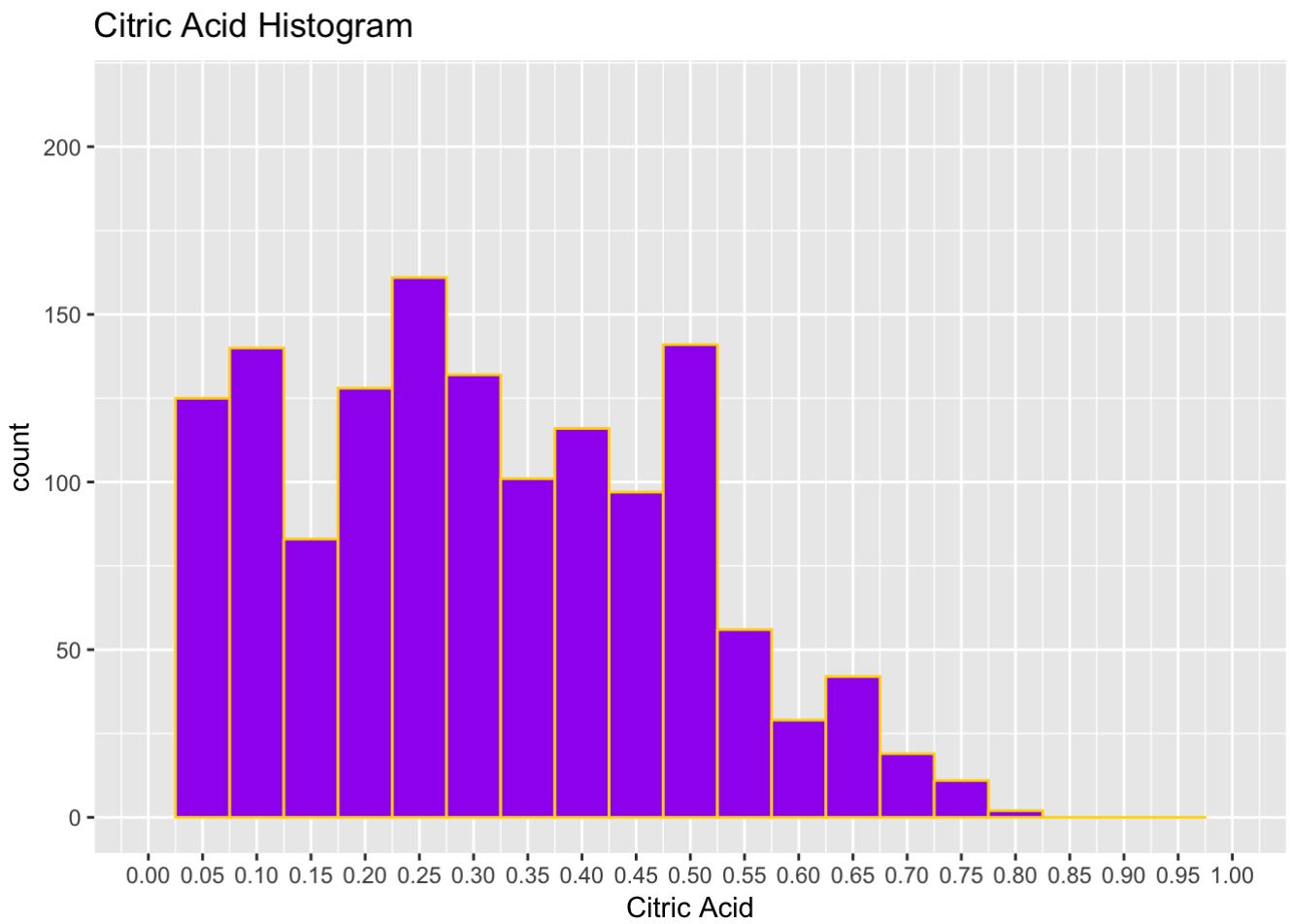
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

```
## Warning: Removed 21 rows containing non-finite values (stat_bin).
```



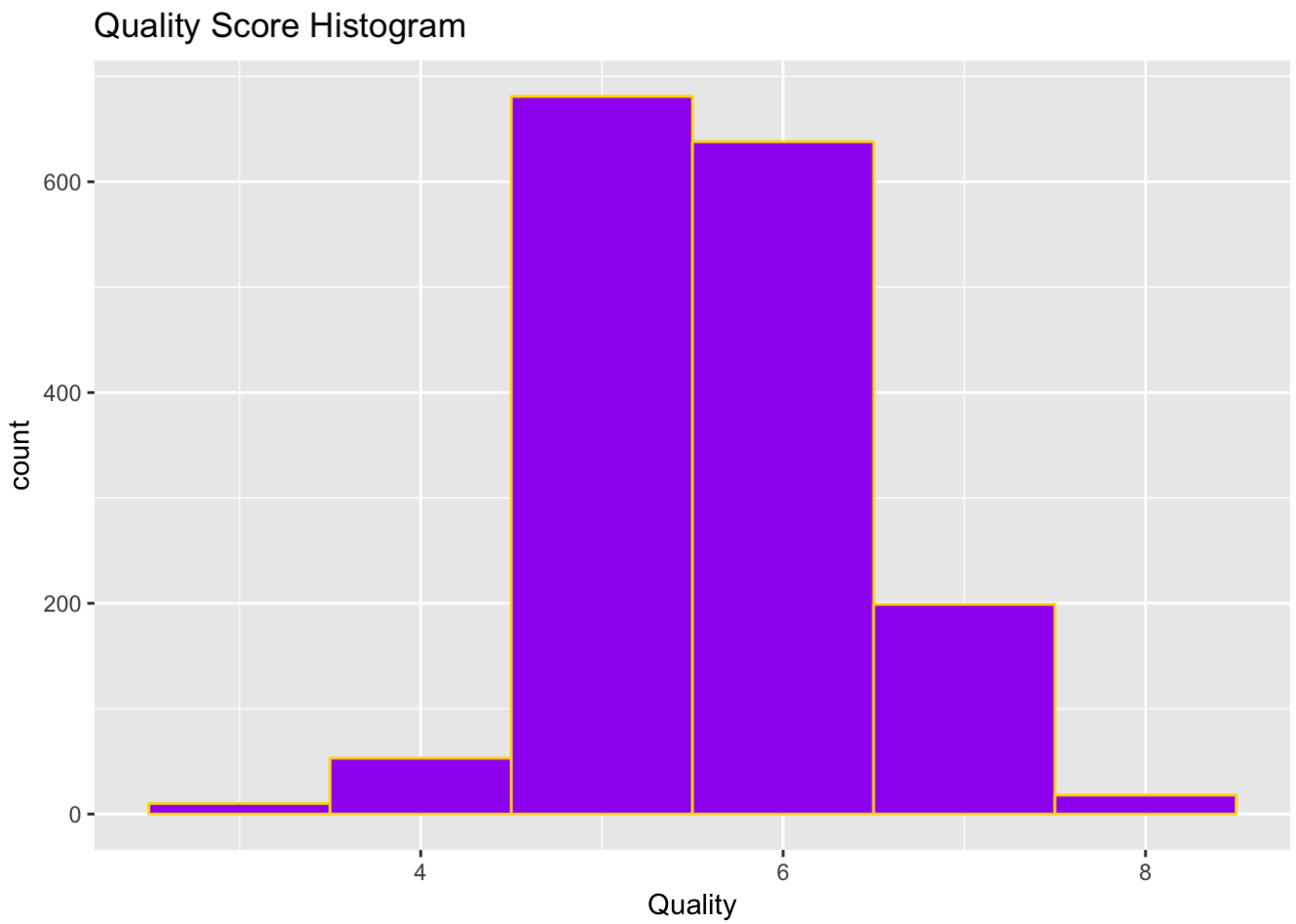
Volatile Acidity has a mean of .5278 and a median of .5200. The data is spread out from 0 to 1 with a sequence of .05. The max Volatile Acidity was 1.58, an outlier, which was taken out of the data set when plotting histogram. Volatile acidity refers to the organic acids found in grape juice, musts and wine that are more volatile or more easily vaporized than the non-volatile or fixed acids

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

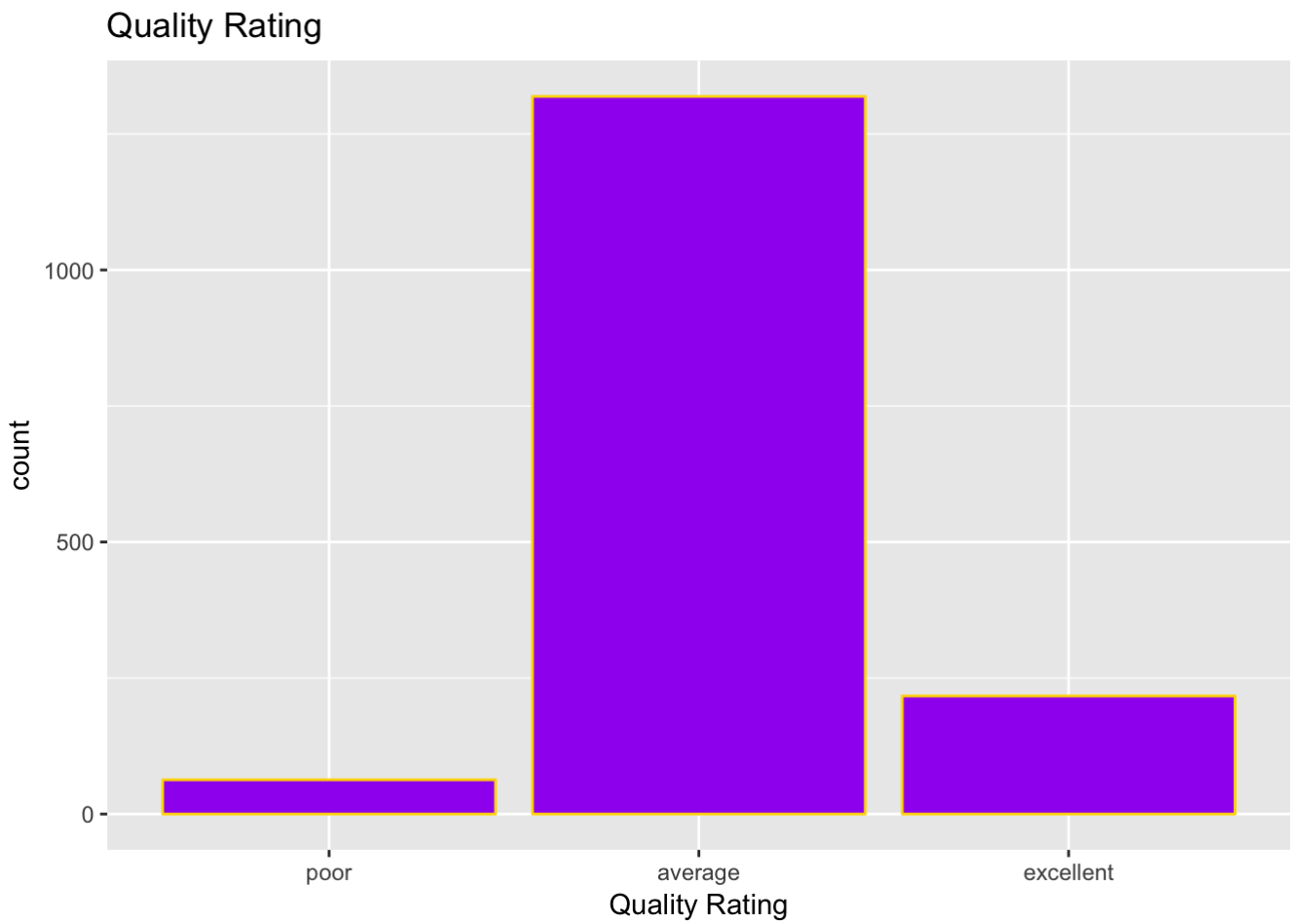


This histogram shows that Citric Acid has regualr peaks and valleys.The average citric acid in the dataset is .271 and the median .271. Citric acid adds sweetness to red wine.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

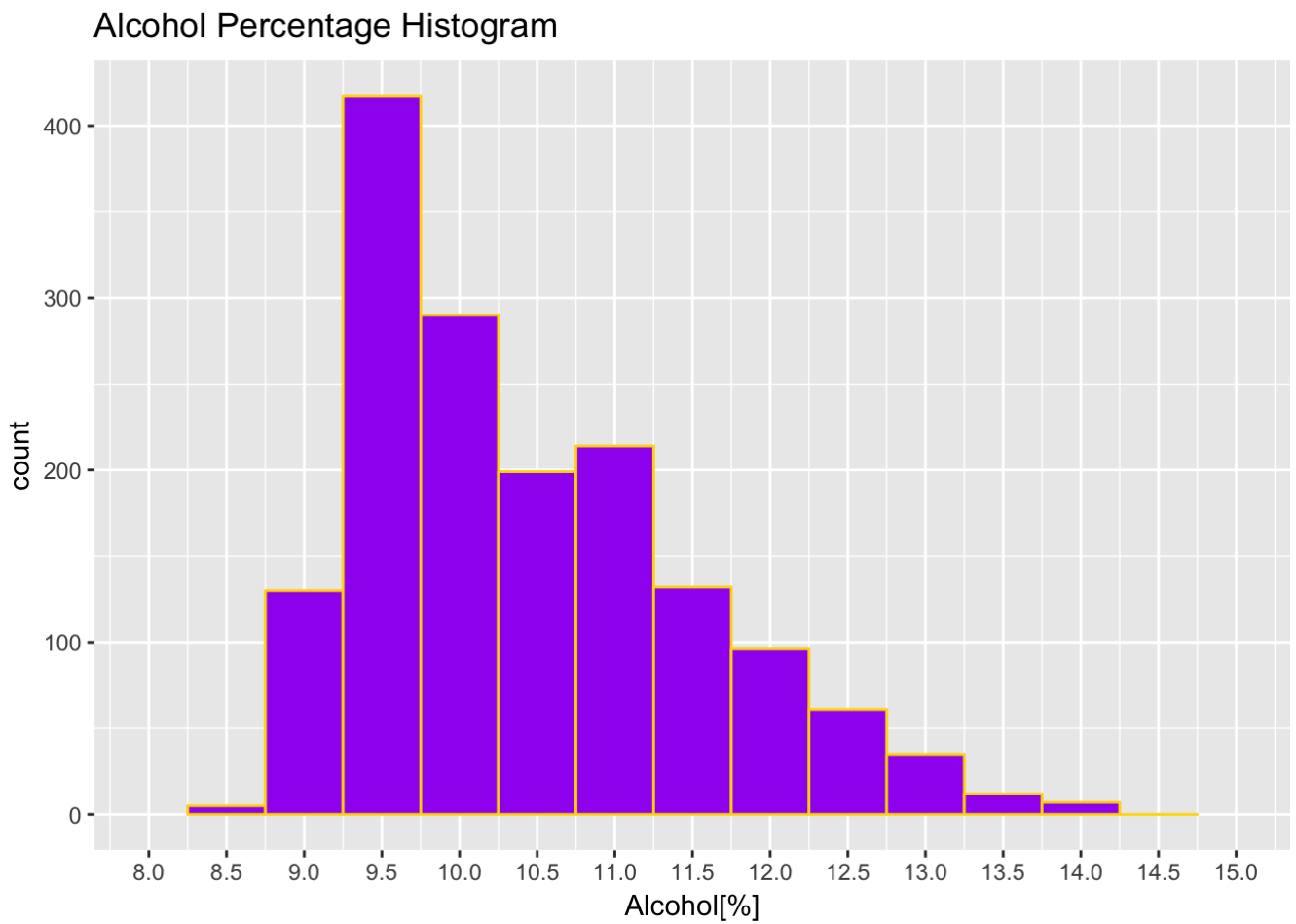



```
##      poor   average excellent
##        63      1319       217
```



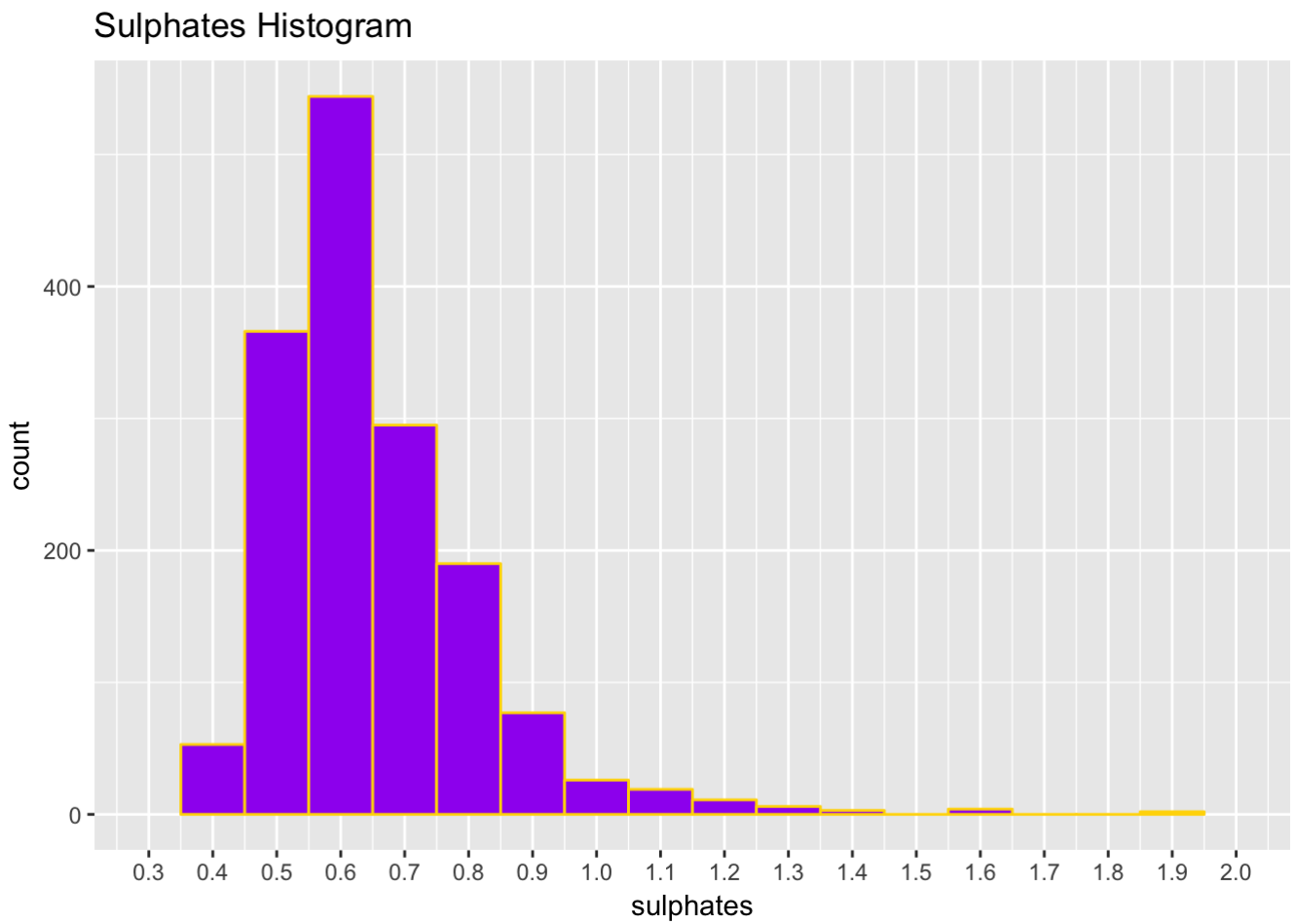
The average quality score for red wine is 5.64 and the median is 6. Scores 5 and 6 are dominated the wine quality scores. 5 and 6 are considered average scores and dominate the histogram. Most wine judgers gave an average score to the red wine recorded in this data set.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```



The majority of red wine in the dataset have an alcohol % of 9.5%. The average amount of alcohol in red wine is 10.42. The histogram seems distributed to the rightly skewed.

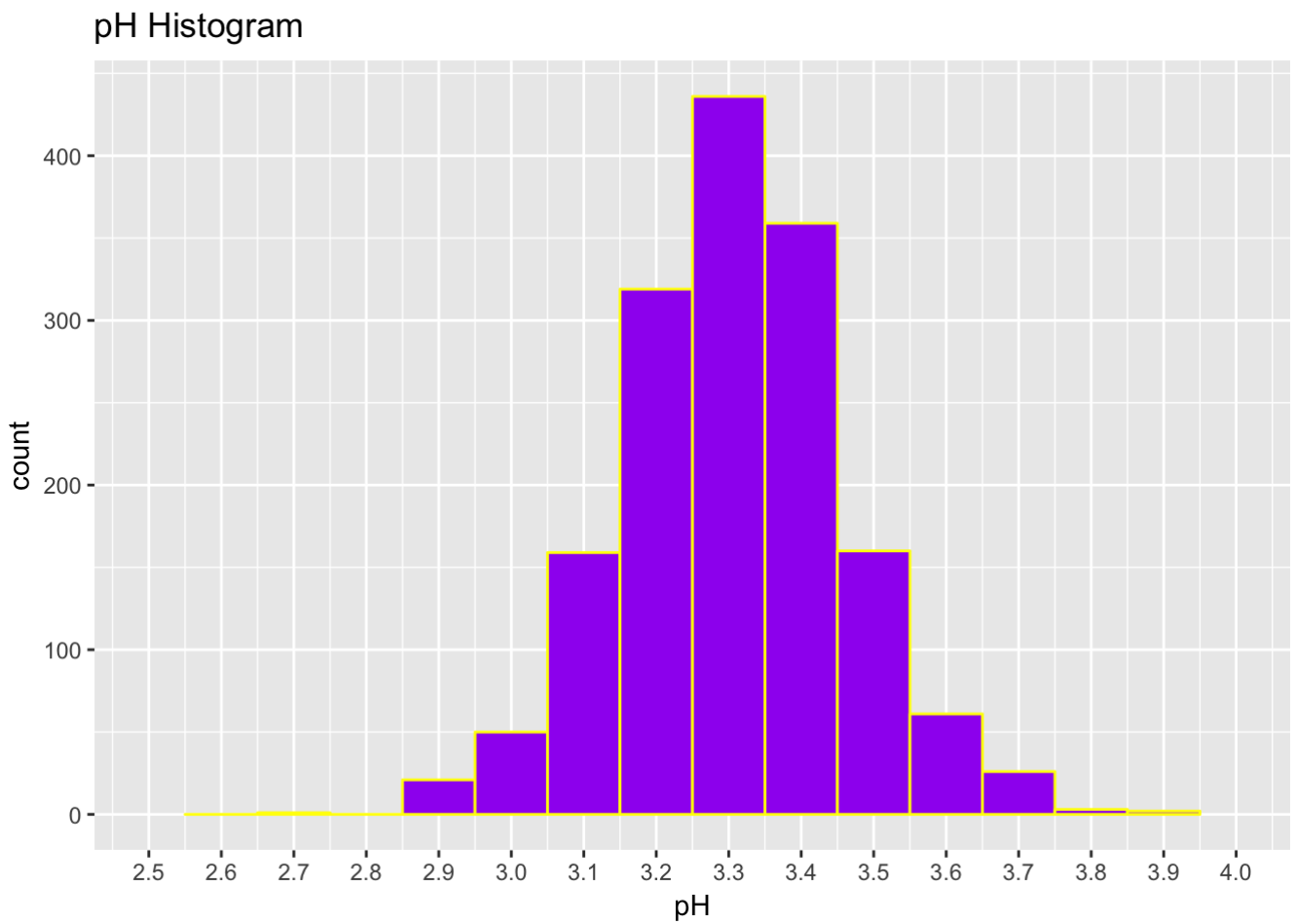
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```



Rightly skewed histogram. The average amount of sulphates according to the summary is .6581 while the median is .6200.The term sulfites is an inclusive term for sulfur dioxide (SO2), a preservative that's widely used in winemaking (and most food industries) for its antioxidant and antibacterial properties.

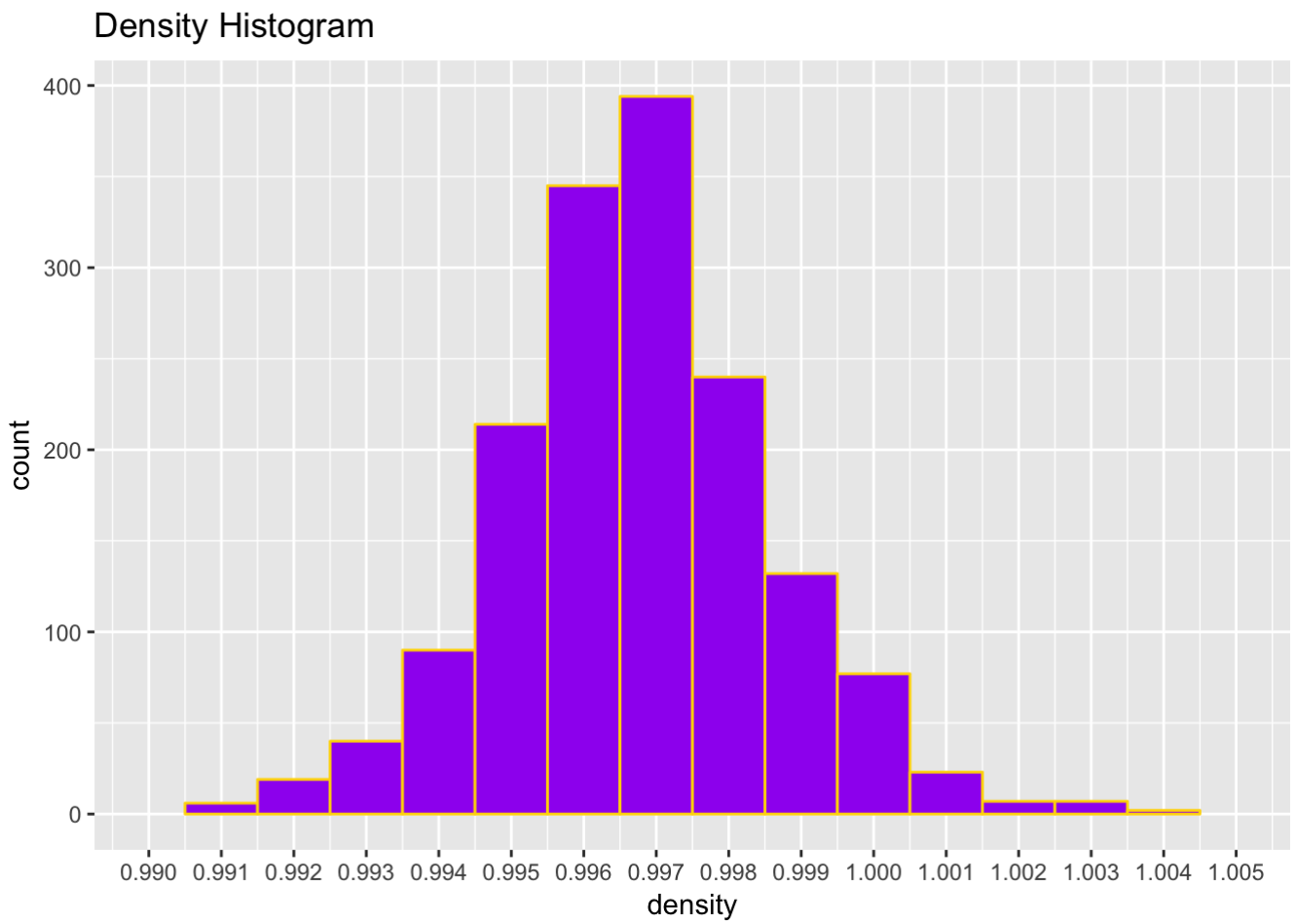
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



pH seems normally distributed. The average for pH is 3.311 while the median is 3.310.

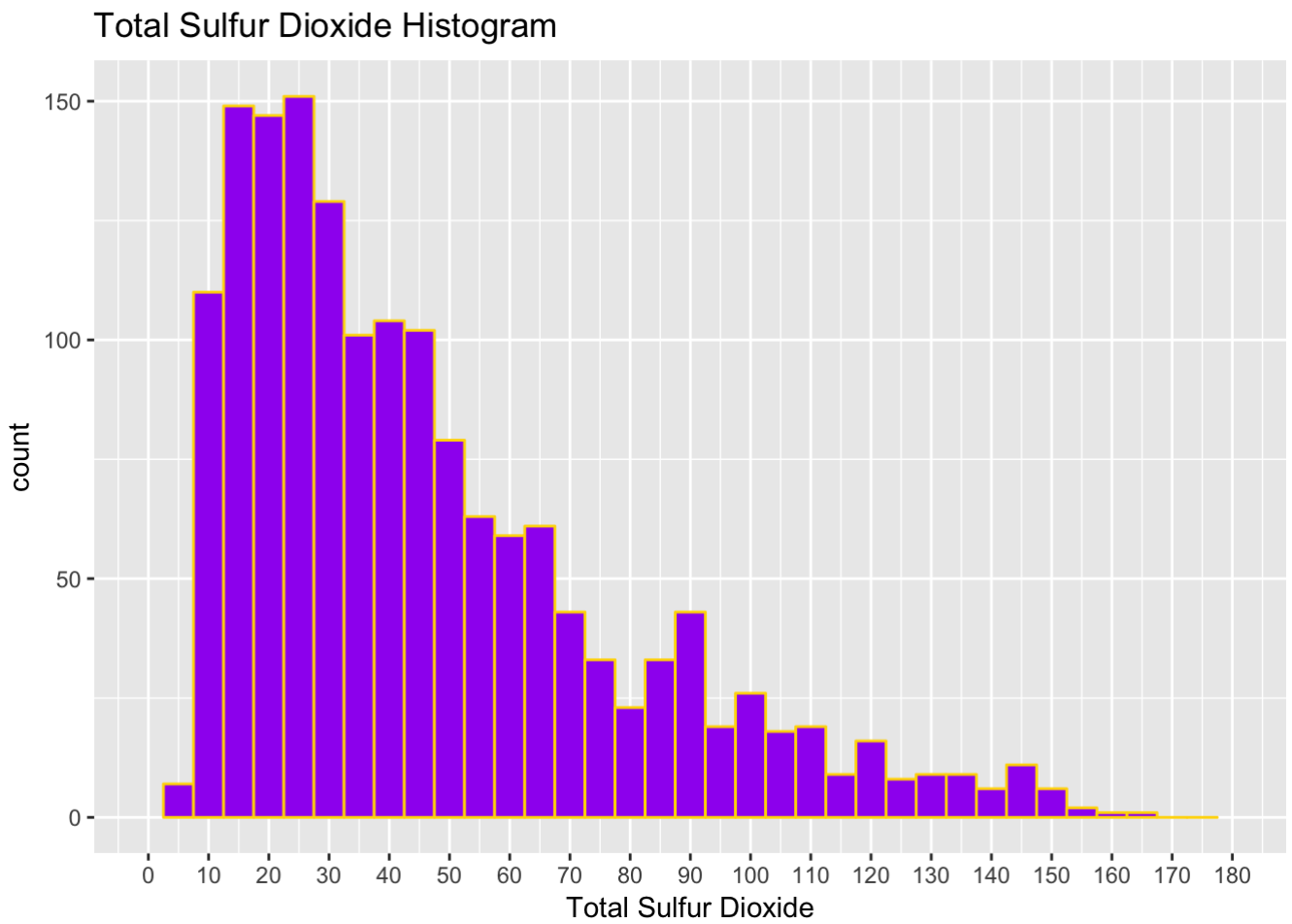
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```



Normally distributed histogram. The mean for density is .9967 and the median is .9968.Density of wine is a measure of the conversion of sugar to alcohol. The must, with sugar but no alcohol, has a high density.

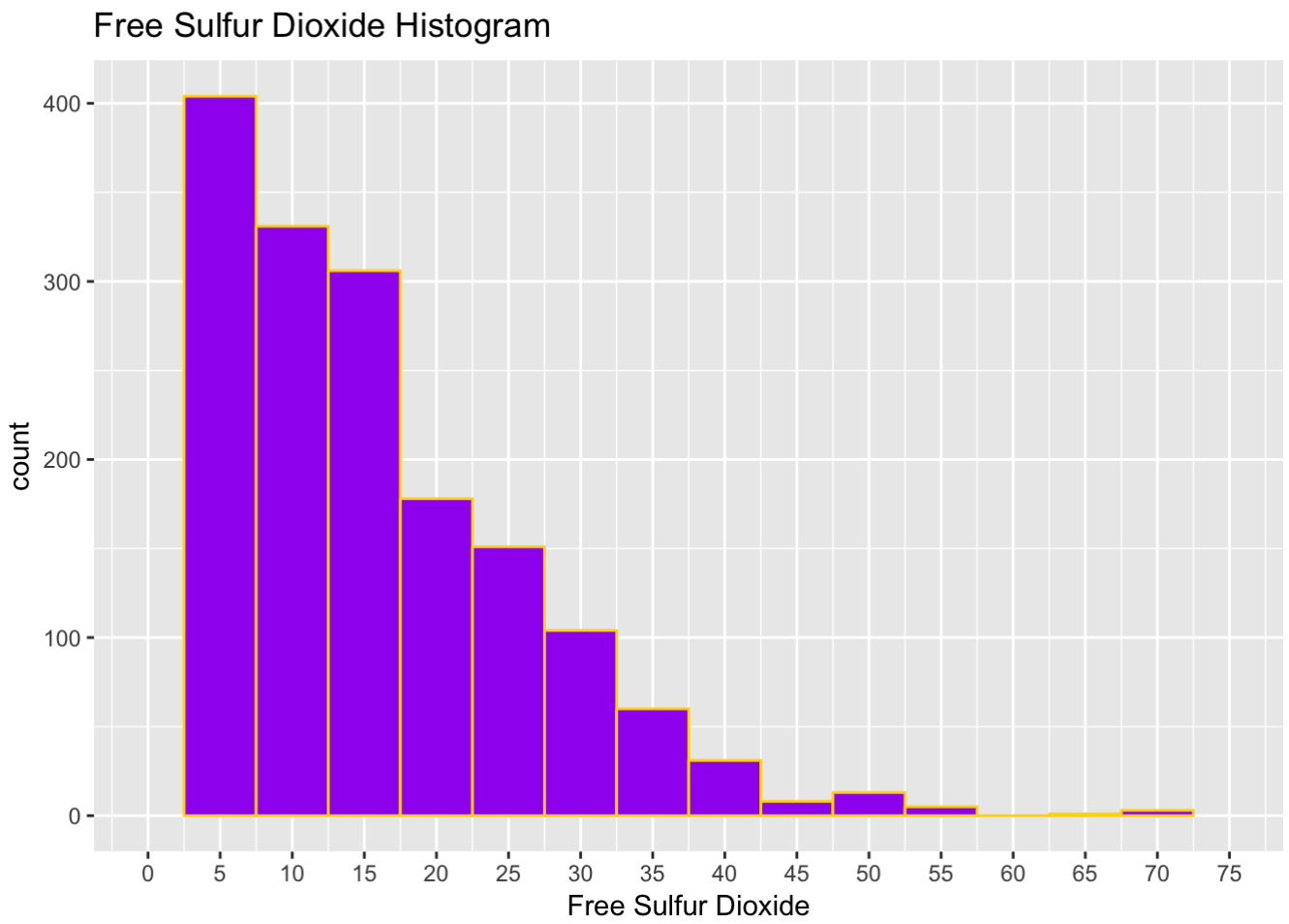
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```



Rightly skewed histogram. Total sulfur dioxide had a mean of 46.67 but a median of 38. The max total sulfur dioxide measure was 289, an outlier. Sulfur dioxide (SO2) is important in the winemaking process as it aids in preventing microbial growth and the oxidation of wine.

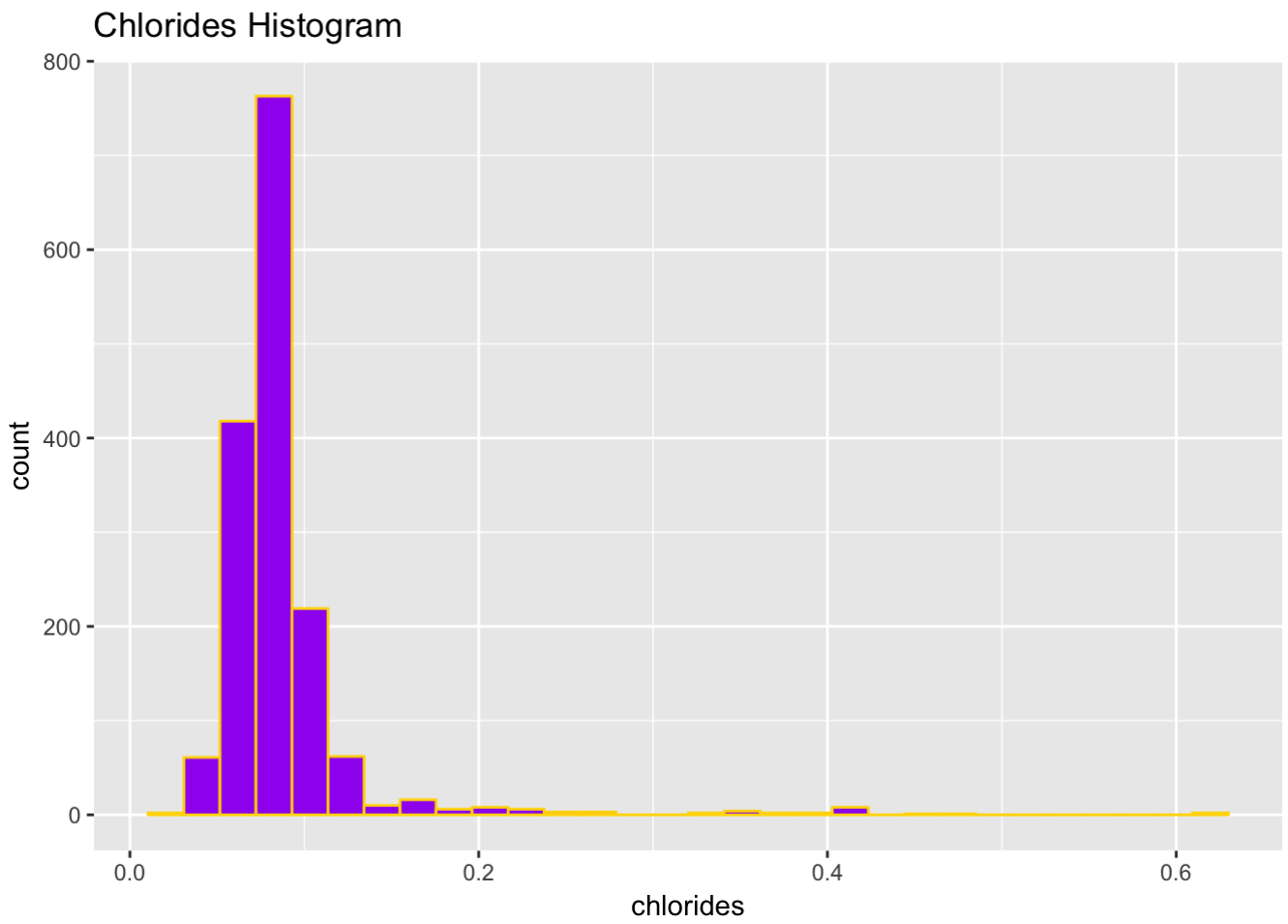
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```



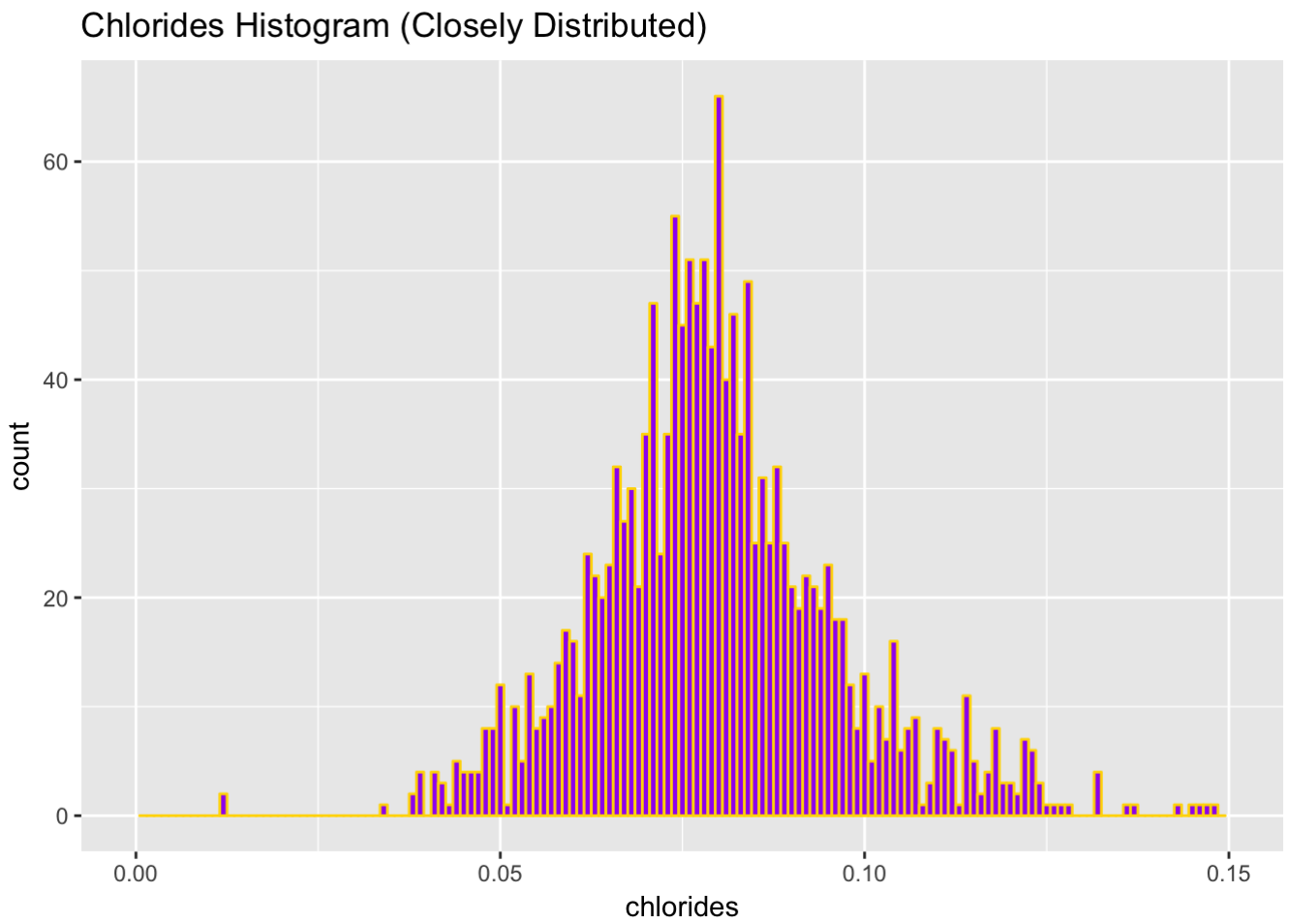
Rightly skewed histogram. Free sulfur dioxide had a mean of 15.87 but a median of 14. The max free sulfur dioxide measure was 72, an outlier. Sulfur dioxide (SO2) is important in the winemaking process as it aids in preventing microbial growth and the oxidation of wine.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



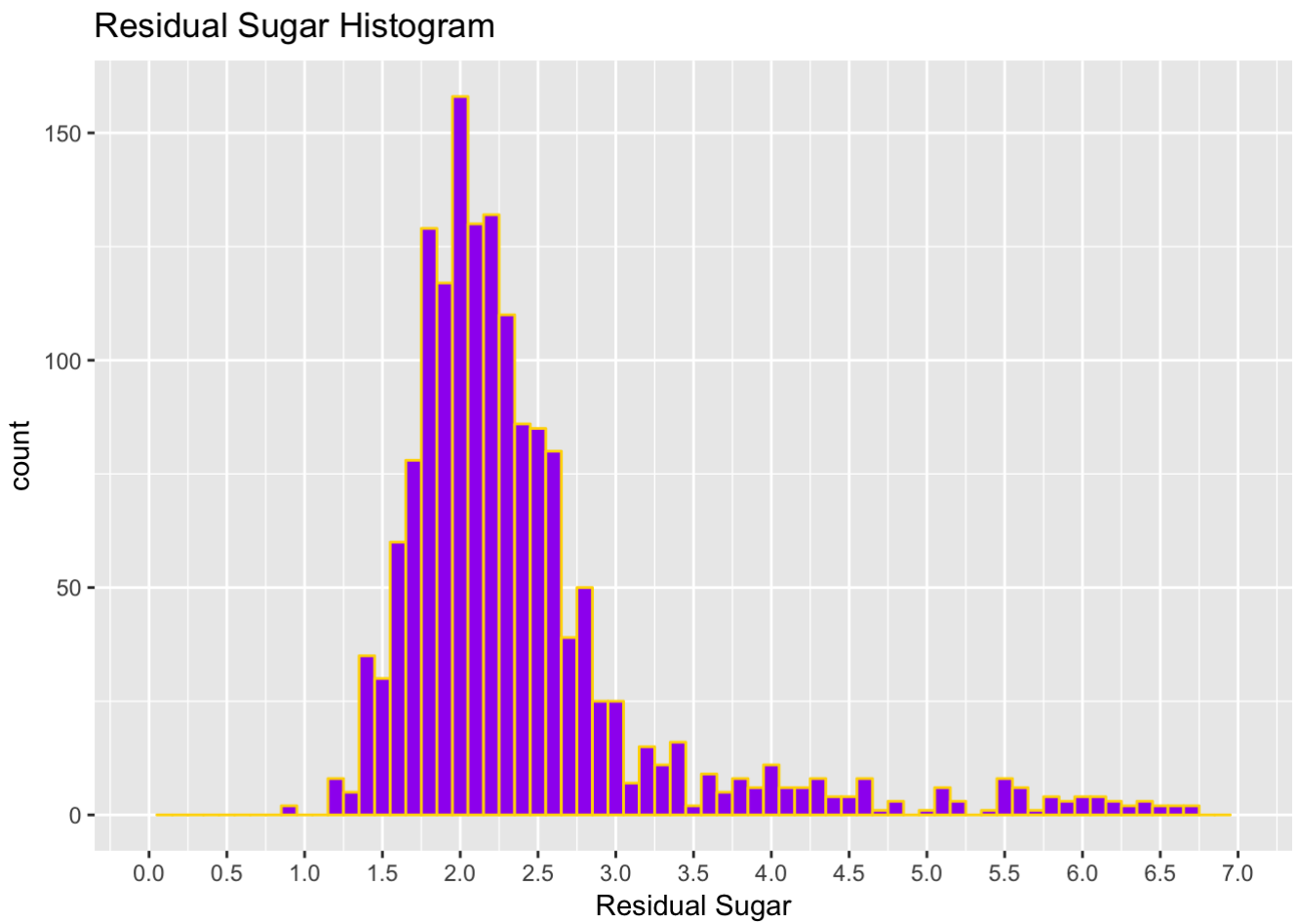

```
## Warning: Removed 67 rows containing non-finite values (stat_bin).
```



Chlorides had a smaller median than average. It had a median of .07900 and a mean of .08747. I decided to use scale_x_continuous to limit the axis from 0 to .15, which shows them more normally distributed.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

```
## Warning: Removed 29 rows containing non-finite values (stat_bin).
```



Residual sugar seems rightly skewed with a median of 2.2 and a mean of 2.539. The max measure for this variable is 15.5, an outlier. The amount of residual sugar refers to any natural grape sugars that are leftover after fermentation ceases.

Univariate Analysis

What is the structure of your dataset?

My dataset has 1599 observations and 13 variables. The variable X is actually a unique identifier in the dataset instead of a measure for red wine. There are 12 measurement variables overall: **residual.sugar,density,alcohol,chlorides,free.sulfur.dioxide,pH,sulphates,quality, citric.acid,quality, fixed.acidity, and volatile.acidity**

Other Observations: The best quality ranges are 7 and above. Most quality ratings fall around 5 or 6 according to the histogram.

Alcohol content tends to be on the lower end of the disstribution as most have a percentage of 9.5%

Chlorides are closely distributed.

Total.sulfur.dioxide has a median of 38 and a mean of 46.67

What is/are the main feature(s) of interest in your dataset?

The main feature of this dataset will be to discover what variables drive quality in red wine.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

The correlation between variables. Even though correlation doesn't mean causation, It would be very interesting to know the tendencies between the red wine measurments and the overall quality of red wine in the dataset.

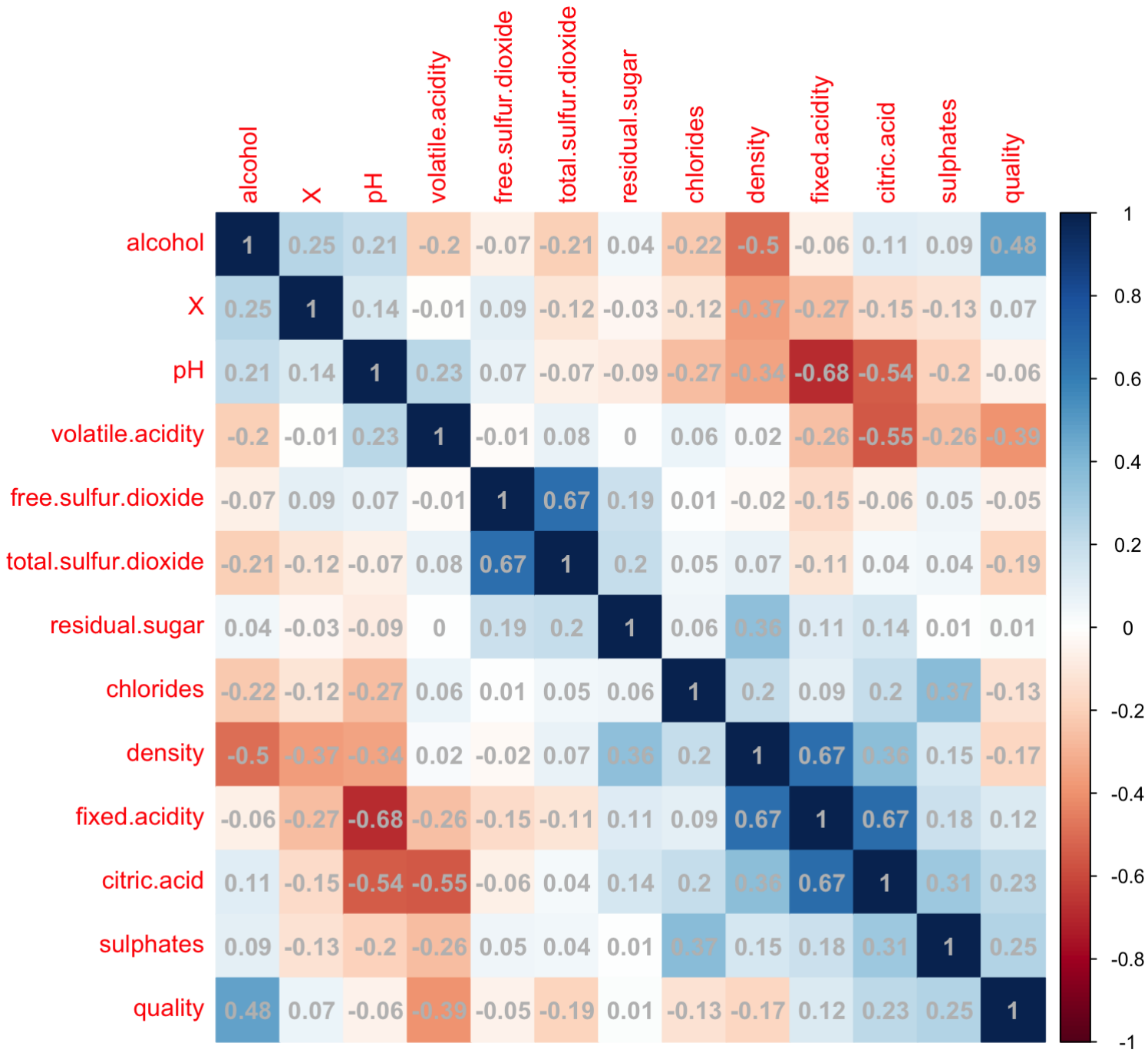
Did you create any new variables from existing variables in the dataset?

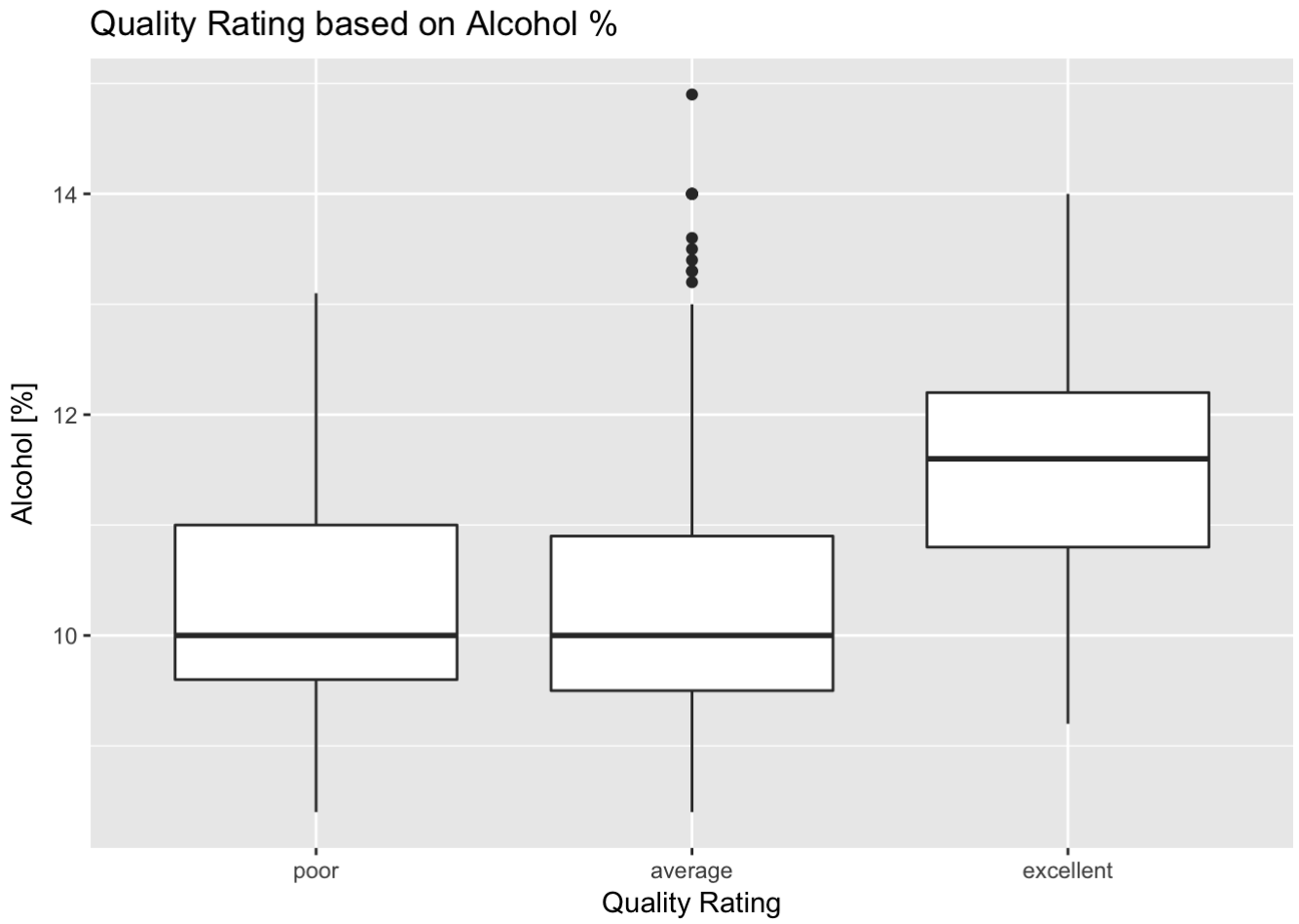
I created a new value called q_rating, which I eventually added to the dataset redInfo. Q_rating is basically a categorical subquality for all of quality entries of red wine from 1 through 8. A score lower than 5 is considered poor. A score between 5 and 6 is considered average. A score of 7 and above is considered excellent.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

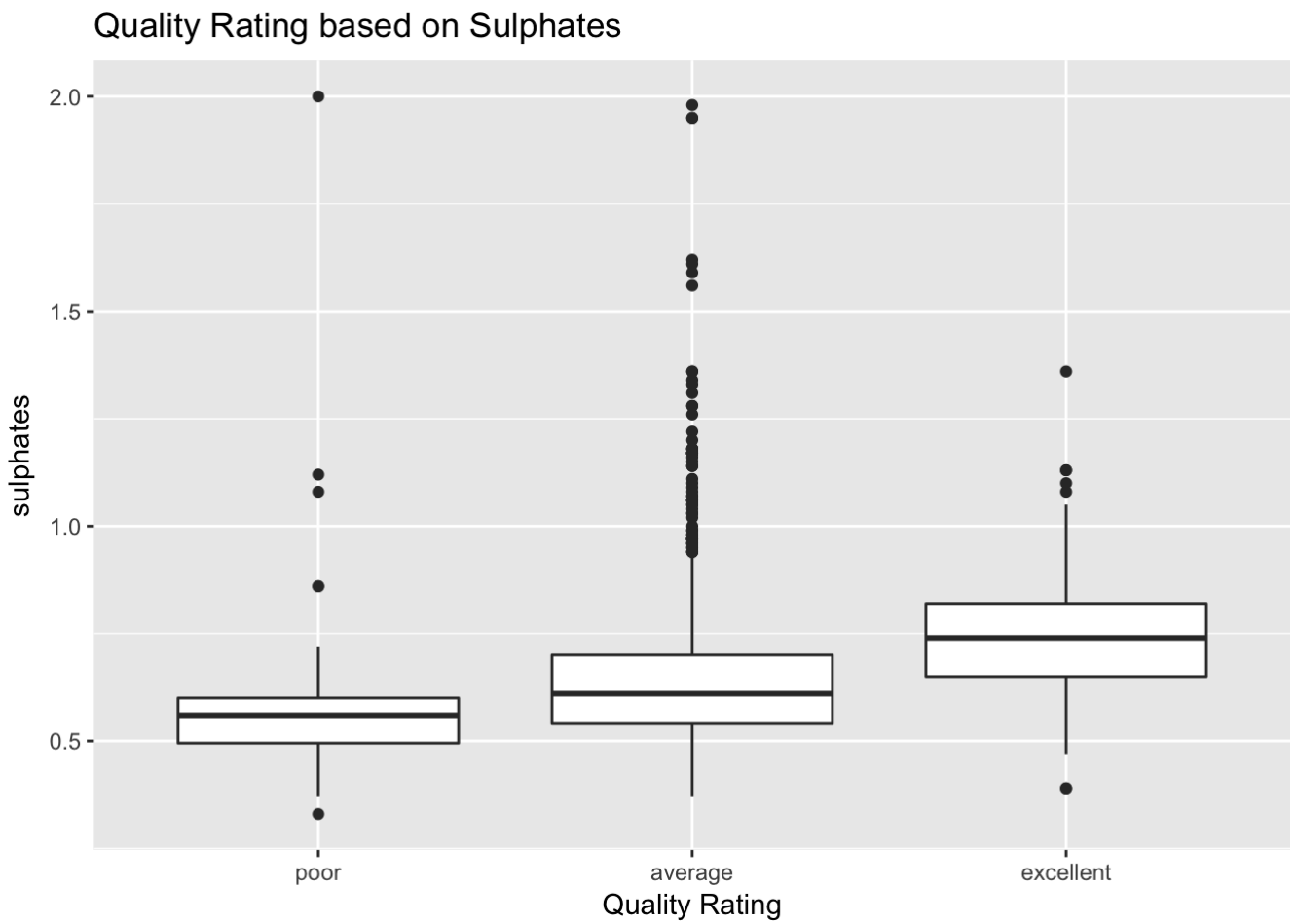
I did change the form of that data by using scale_x_contiunious and changing to the appropriate binwidths based on the variable. It was used to get a better look of the distribution as well as getting rid of outliers. Most variables were rightly skewed.

Bivariate Plots Section

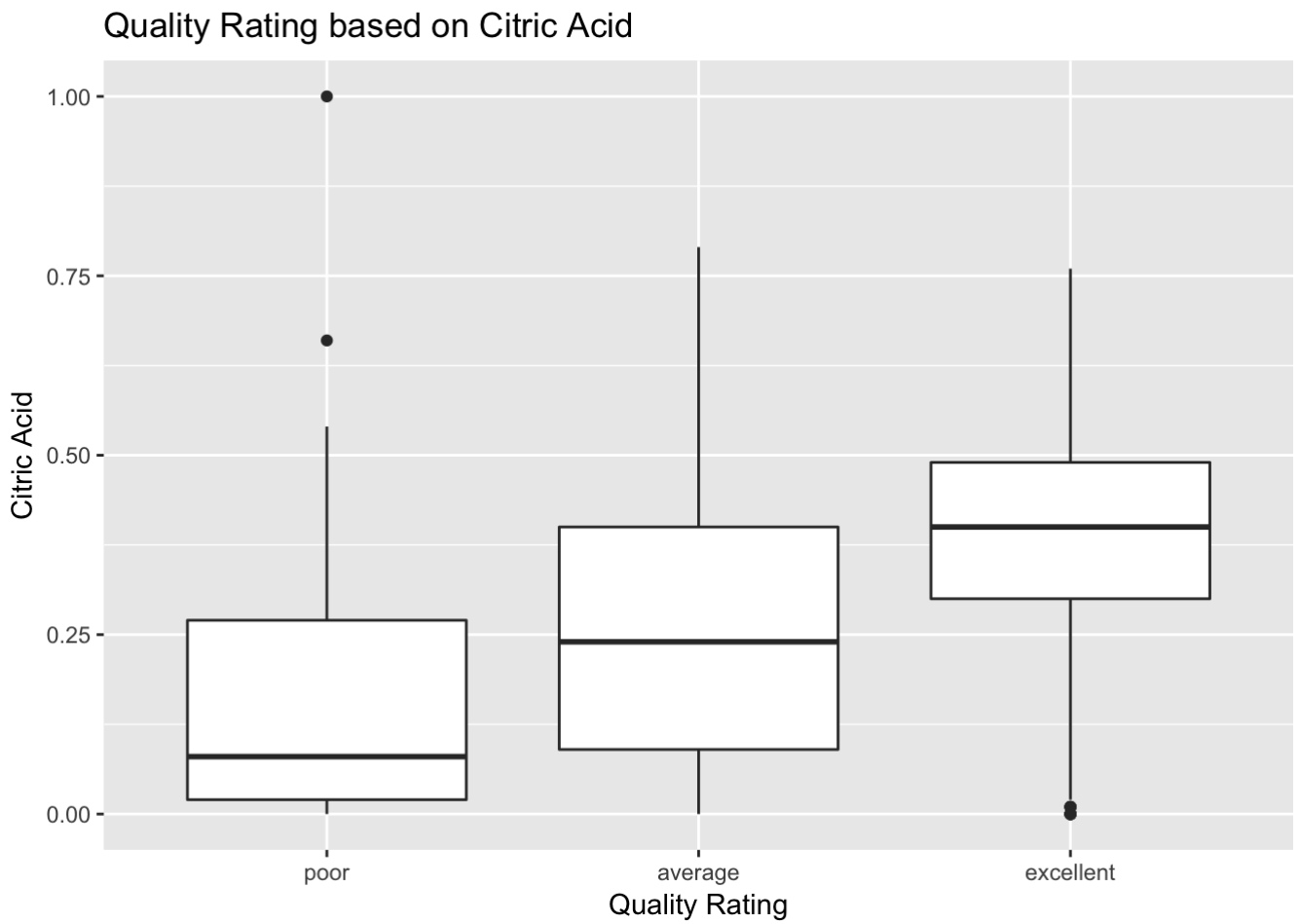

alcohol
X
pH
volatile.acidity
free.sulfur.dioxide
total.sulfur.dioxide
residual.sugar
chlorides
density
fixed.acidity
citric.acid
sulphates
quality
alcohol 1 0.25 0.21 -0.2 -0.07 -0.21 0.04 -0.22 -0.5 -0.06 0.11 0.09 0.48
X 0.25 1 0.14 -0.01 0.09 -0.12 -0.03 -0.12 -0.37 -0.27 -0.15 -0.13 0.07
pH 0.21 0.14 1 0.23 0.07 -0.07 -0.09 -0.27 -0.34 -0.68 -0.54 -0.2 -0.06
volatile.acidity -0.2 -0.01 0.23 1 -0.01 0.08 0 0.06 0.02 -0.26 -0.55 -0.26 -0.39
free.sulfur.dioxide -0.07 0.09 0.07 -0.01 1 0.67 0.19 0.01 -0.02 -0.15 -0.06 0.05 -0.05
total.sulfur.dioxide -0.21 -0.12 -0.07 0.08 0.67 1 0.2 0.05 0.07 -0.11 0.04 0.04 -0.19
residual.sugar 0.04 -0.03 -0.09 0 0.19 0.2 1 0.06 0.36 0.11 0.14 0.01 0.01
chlorides -0.22 -0.12 -0.27 0.06 0.01 0.05 0.06 1 0.2 0.09 0.2 0.37 -0.13
density -0.5 -0.37 -0.34 0.02 -0.02 0.07 0.36 0.2 1 0.67 0.36 0.15 -0.17
fixed.acidity -0.06 -0.27 -0.68 -0.26 -0.15 -0.11 0.11 0.09 0.67 1 0.67 0.18 0.12
citric.acid 0.11 -0.15 -0.54 -0.55 -0.06 0.04 0.14 0.2 0.36 0.67 1 0.31 0.23
sulphates 0.09 -0.13 -0.2 -0.26 0.05 0.04 0.01 0.37 0.15 0.18 0.31 1 0.25
quality 0.48 0.07 -0.06 -0.39 -0.05 -0.19 0.01 -0.13 -0.17 0.12 0.23 0.25 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



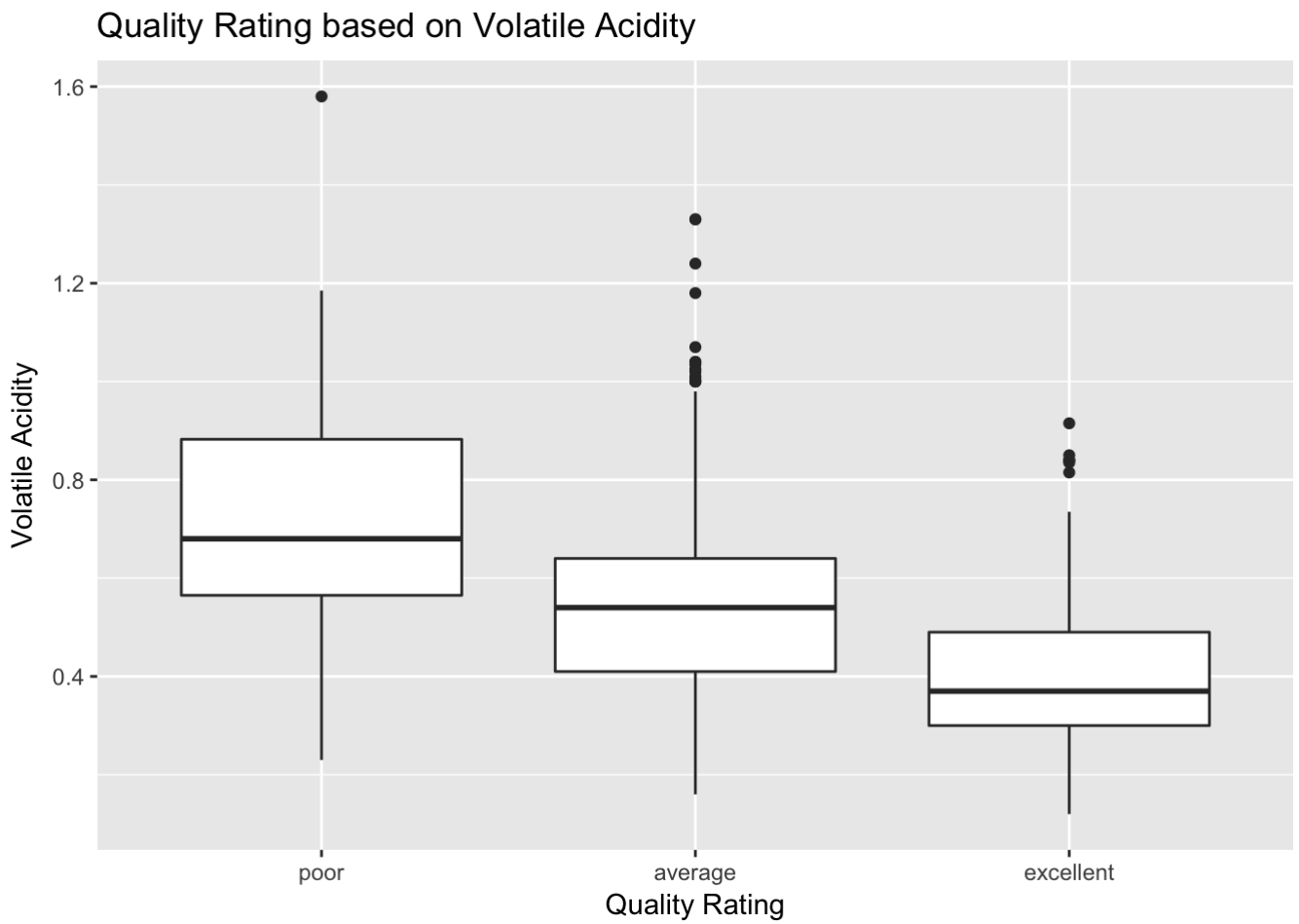

```
## q_rating: poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.60   10.00   10.22   11.00   13.10
## ------------------------------------------------------------
## q_rating: average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.00   10.25   10.90   14.90
## ------------------------------------------------------------
## q_rating: excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.60   11.52   12.20   14.00
```



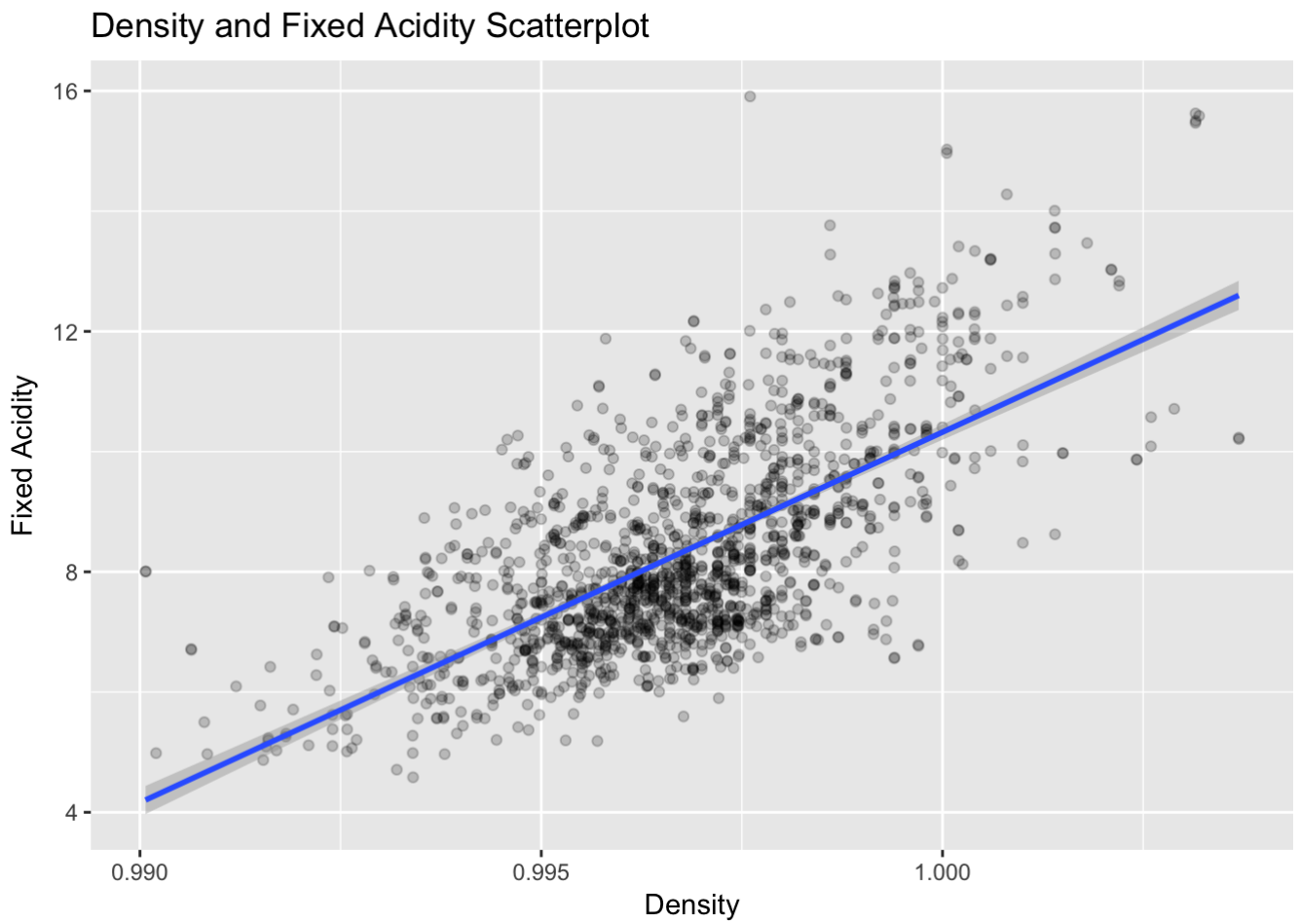

```
## q_rating: poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.4950  0.5600  0.5922  0.6000  2.0000
## ------------------------------------------------------------
## q_rating: average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3700  0.5400  0.6100  0.6473  0.7000  1.9800
## ------------------------------------------------------------
## q_rating: excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3900  0.6500  0.7400  0.7435  0.8200  1.3600
```



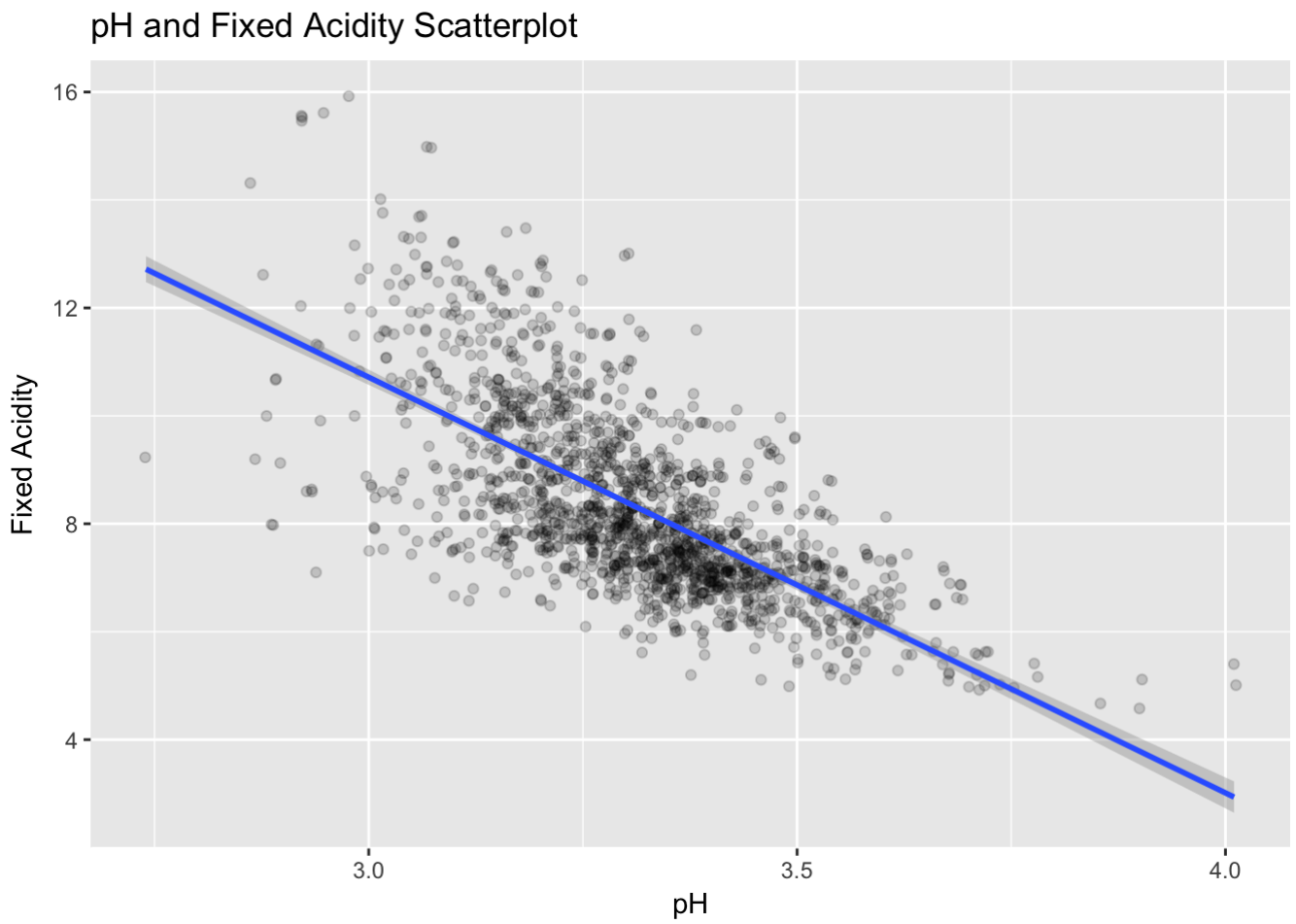

```
## q_rating: poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0200  0.0800  0.1737  0.2700  1.0000
## ------------------------------------------------------------
## q_rating: average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2400  0.2583  0.4000  0.7900
## ------------------------------------------------------------
## q_rating: excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3000  0.4000  0.3765  0.4900  0.7600
```



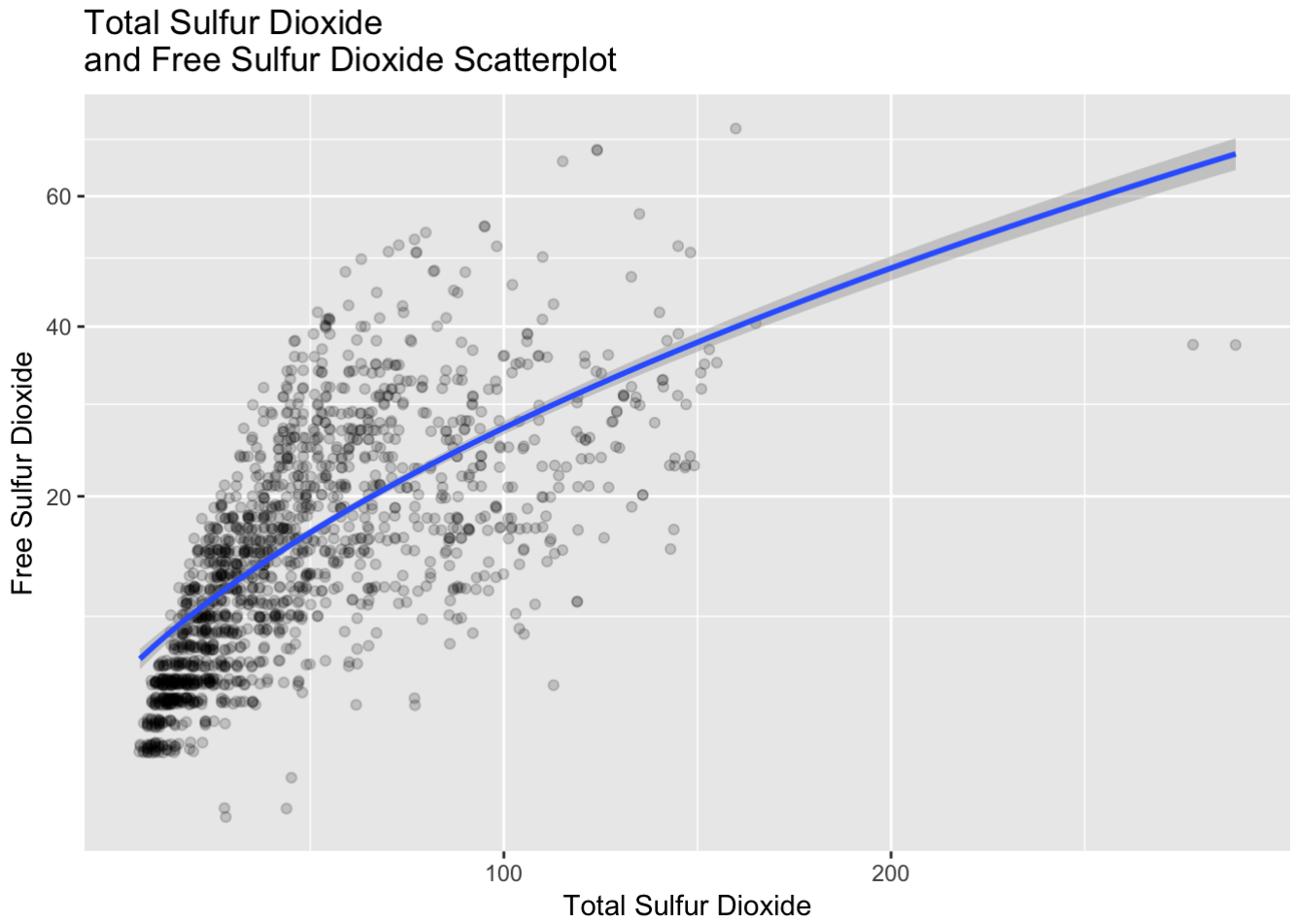

```
## q_rating: poor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2300  0.5650  0.6800  0.7242  0.8825  1.5800
## ------------------------------------------------------
## q_rating: average
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.4100  0.5400  0.5386  0.6400  1.3300
## ------------------------------------------------------
## q_rating: excellent
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4055  0.4900  0.9150
```


Density and Fixed Acidity Scatterplot
16
12
8
4
Fixed Acidity
0.990
0.995
1.000
Density

```
## [1] 0.6680473
```


pH and Fixed Acidity Scatterplot
16
12
8
4
Fixed Acidity
3.0
3.5
4.0
pH

```
## [1] -0.6829782
```




```
## [1] 0.6676665
```

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

My feature of interest is determing what are the main factors that contribute to the quality of wine in the dataset.It seems that the wine has 4 variables that play a very considerable role in terms of having a good quality soce. Those variable are : alcohol, volatile.acidity, citric.acid, and sulphates.

As someone who doesn't drink or have any initial knowledge about wine, I wasn't expecting anything particularly. Before plotting anything, I decided to make a correlation plot between all variables. I believed that it would give me some sense of direction to analyze and explore them. The charts, referenced by the variable pairings below, represents the variables that had high correlations, specifically for the variable quality.

Since quality was the most important factor, it was important that I find and analyze the variables that had a considerably high corelation with it. 4 variables stood out: Alcohol, citric acid, volatile acidity, and sulphates. Other variable pairings that didn't involve quality were considered eye opening as well such as total sulfurdioxide and free sulfur dioxide , pH and fixed acidity, and density and fixed acidity. Even though these variable correlation pairing were not my area of focus, I believed that it could be very helpful in the multivariate section. Below is a description of the variable pairings that had a considerably high correlation with the variable quality.

Quality Rating and Alcohol There is a .48 correlation between alcohol and quality rating. According the the boxplot, red wines with more alcohol content tend to have better quality scores. For poor and average scores, the median for alcohol is 10.00, but for excellent scores the median is 11.60. The mean for alcohol in excellent scores is 11.52, which is higher than the mean for poor, average, and total average quality scores.

Quality Rating and Volatile Acidity There is a -0.39 correlation between volatile acidity and quality rating. Accroding to the boxplot, the higher the quality rating, the less volatile acidity the red wine has.Too much volatile acid can turn the wine into vinegar (or give it a vinegar-like taste). Poor rated wines had the highest median and mean scores of .68 and .724 respectively. Average rated wines have a the second highest median and mean scores of .54 and .5386 respectively. The excellent rated red wines, had the lowest and best median and mean scores of .37 and .4055.

Quality Rating and Sulphates There is a .25 correlation between quality rating and sulphates. According to the boxplot, better quality ratings tend to have more sulphates in the red wine.Sulphates (sulfur dioxide (SO2)), is a preservative that's widely used in winemaking (and most food industries) for its antioxidant and antibacterial properties.The excellent quality red wines had median and mean scores of .74 and .7435 respectively while the poor quality red wines had scores of .56 and .5922.

Quality Rating and Citric Acid There is a .23 correlation between quality rating and citric acid. According the the boxplot, better quality rattings tend to have more citric acid in the red wine.Citric acid is produced with the help of special mold and used as a flavor enhancer and preservative in food. The excellent quality red wines had median and mean scores of .4 and .3765 respectively while the poor quality red wines had scores of .0800 and .1737.

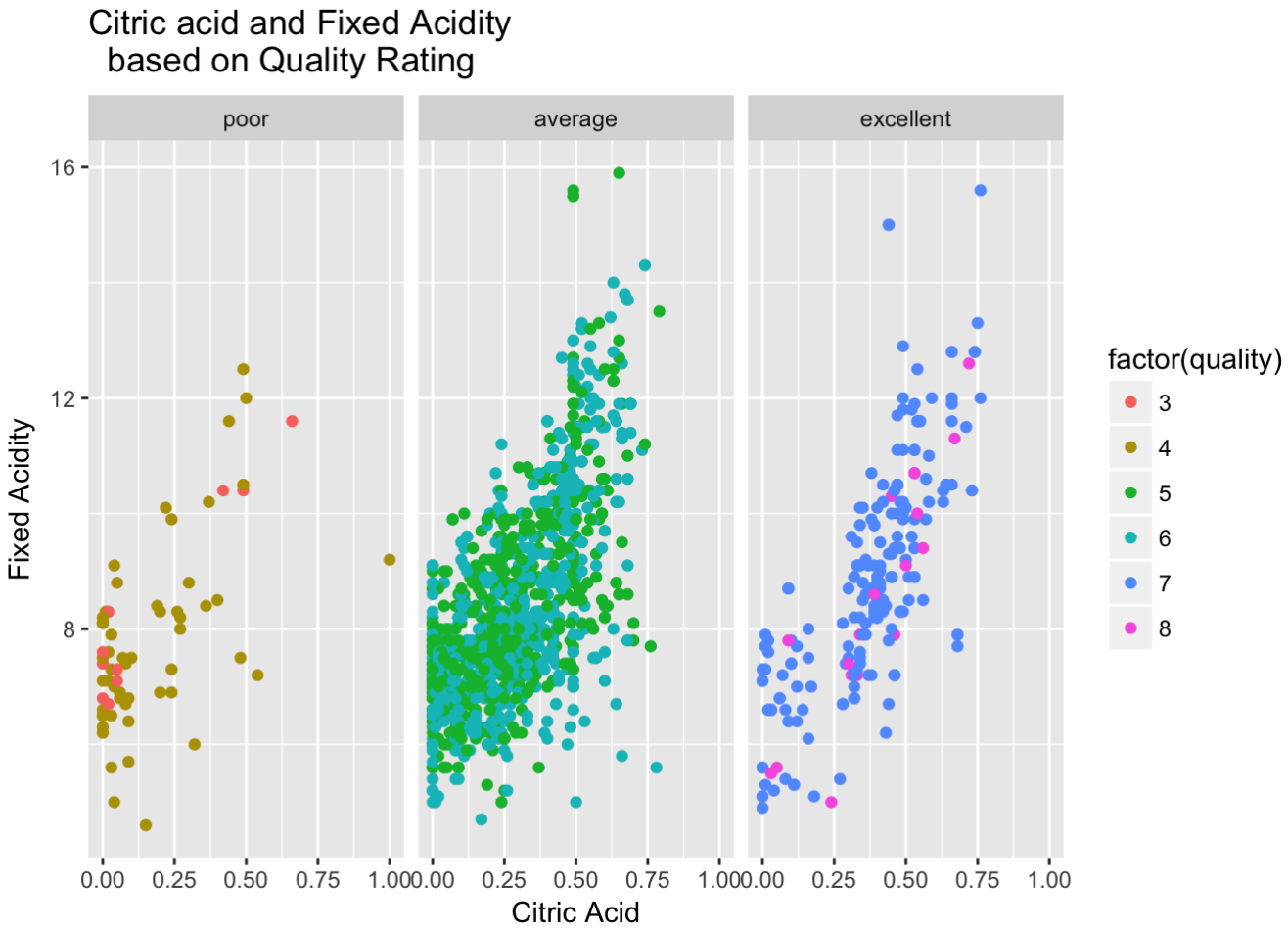
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

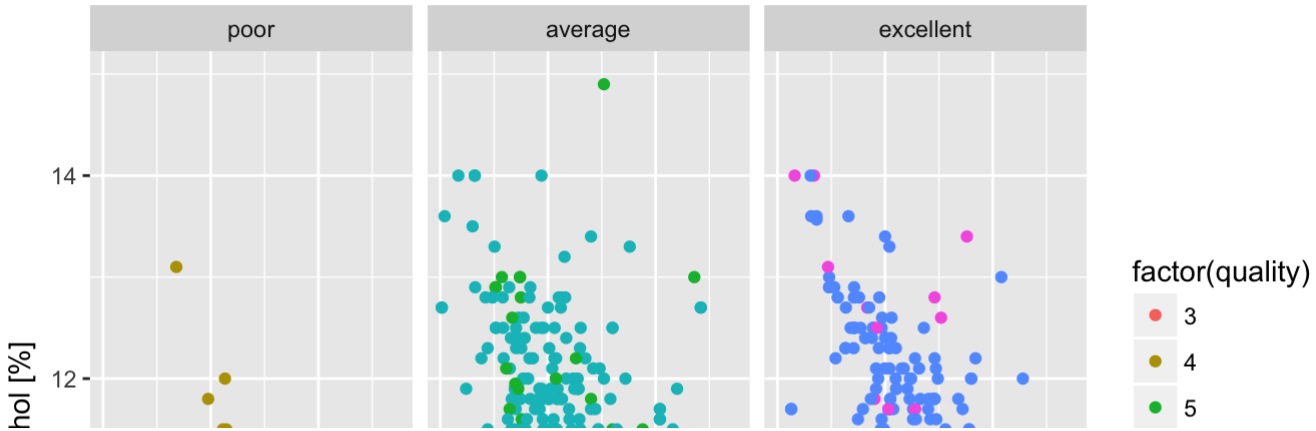
Due to high positive and negative correltions, I also took interest in these correlated variables: total.sulfur.dioxide and free sulfur dioxide (positively correlated with 0.668), pH and fixed acidity (negatively correlated with -0.683), and density and fixed acidity (positively correlated with 0.668).

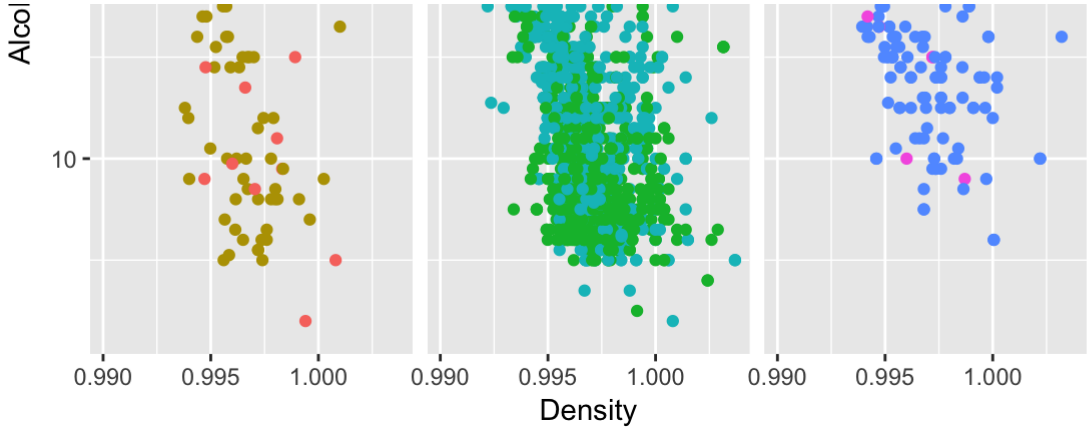
What was the strongest relationship you found?

The strongest relationship was between Fixed Acidity and pH, which had a correlation coefficient of -0.683.

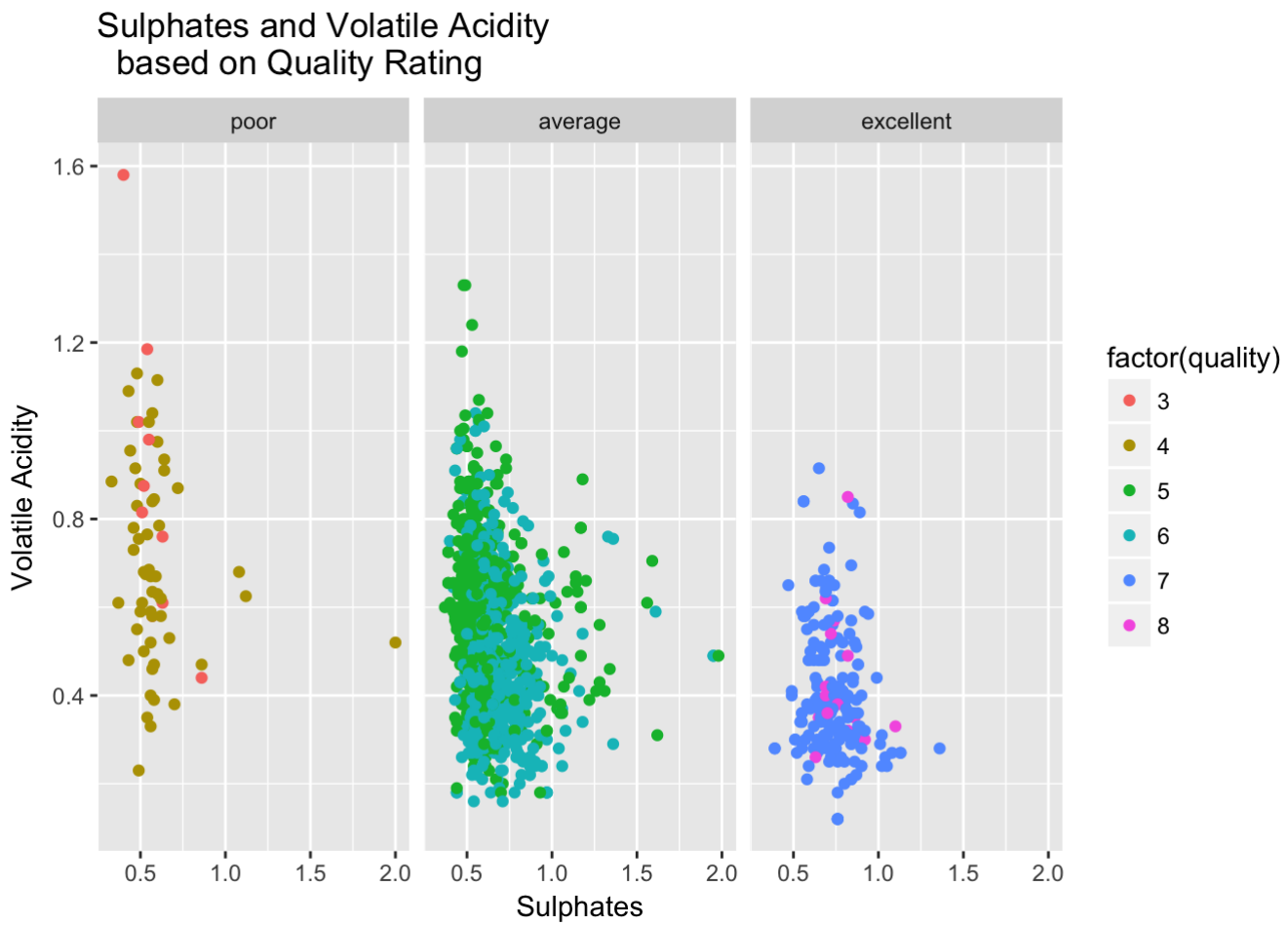
Multivariate Plots Section


Citric acid and Fixed Acidity
based on Quality Rating
poor
average
excellent
Fixed Acidity
16
12
8
0.00
0.25
0.50
0.75
1.00
Citric Acid
factor(quality)
3
4
5
6
7
8


Alcohol % and Density
based on Quality Rating
poor
average
excellent
14
12
factor(quality)
3
4
5


Alcol
10
0.990
0.995
1.000
Density
6
7
8


Citric Acid and Alcohol %
based on Quality Rating
poor
average
excellent
Alcohol [%]
14
12
10
0.00
0.25
0.50
0.75
1.00
Citric Acid
factor(quality)
3
4
5
6
7
8



Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

I wanted to take a closer look at variables from the bivariate section that were the main features of interest due to their relationship with the quality variable and their high correlations.

In the Bivariate section of this project, I learned that quality had a considerable correlation with 4 different variables. The next step was to see how these 4 variables (Citric Acid, Volatile Acidity, Alcohol %, and Sulphates) correlated with each other by looking at the correlation plot in the last section to see their relationship with quality.

I expected to find two pairs of considerable correlations between the two variables and was successful: the pairs of Sulphates and Volatile Acidity, and Citric Acid and Alcohol. There were some other pairs worth exploring that I've found to have high correlations as well: Density and Alcohol as well as Fixed Acidity and Citric Acid.

Fixed Acidity and Citric Acid based on Quality There is a .67 correlation between fixed acidity and citric acid. Both of these variables had considerable correlations individually with quality so I decided to plot them both using the quality rating variable. According to the plot, the red wines with higher citric acid tended to have higher fixed acidity measurables. Excellent wines tend to have citric acid content betwen .30 and .50 and higher concentrations of fixed acidity. Average wines had a median score of 7.8 but did not have high concentrations of fixed acidity. The red wins that did have high concentrations were wines that had a rating of 6.

Density And Alcohol based on Quality There is an obvious relationshion between density and alcohol based on quality. Excellent red wines had the lowest median and average scores in terms of density, but hat the most alcohol content which was expected.Based on the plot, average red wines show a considerable amount of density while also having lower alcohol content. Wine judgers could be biased and would prefer wine that is less dense and has good amount of alcohol content.

Citric Acid and Alcohol based on Quality Alcohol content and citric acid aren't highly correlated but there is a positive relationship between the two variables. poor quality red wines tended to be lower in alcohol content and citric acid. Alcohol content made average wines taste better regardless of citric acid content. Excellent wines tended to be higher in alcohol content and citric acid. A trend is also noticable in this plot as well. poor red wines tend to have low citric acid scores and low alcohol content. From poor to excellent, the trend seems to go upward. Alcohol plays an important role in determining the quality of wine.

Sulphates and Volatile Acidity based on Quality Poor wines tend to have high rates volatile acidity while having a low amount of sulphates. The plot shows a trent. for each subquality, it tends to decrease. Average wines have a lower volatile acidity than poor wines , but have higher rates of sulphates. Out of all subquality ratings, excellent red wines had lower sulphates and low volatile acidity. This would be a revelation for determining what really makes a poor red wine and great wine.

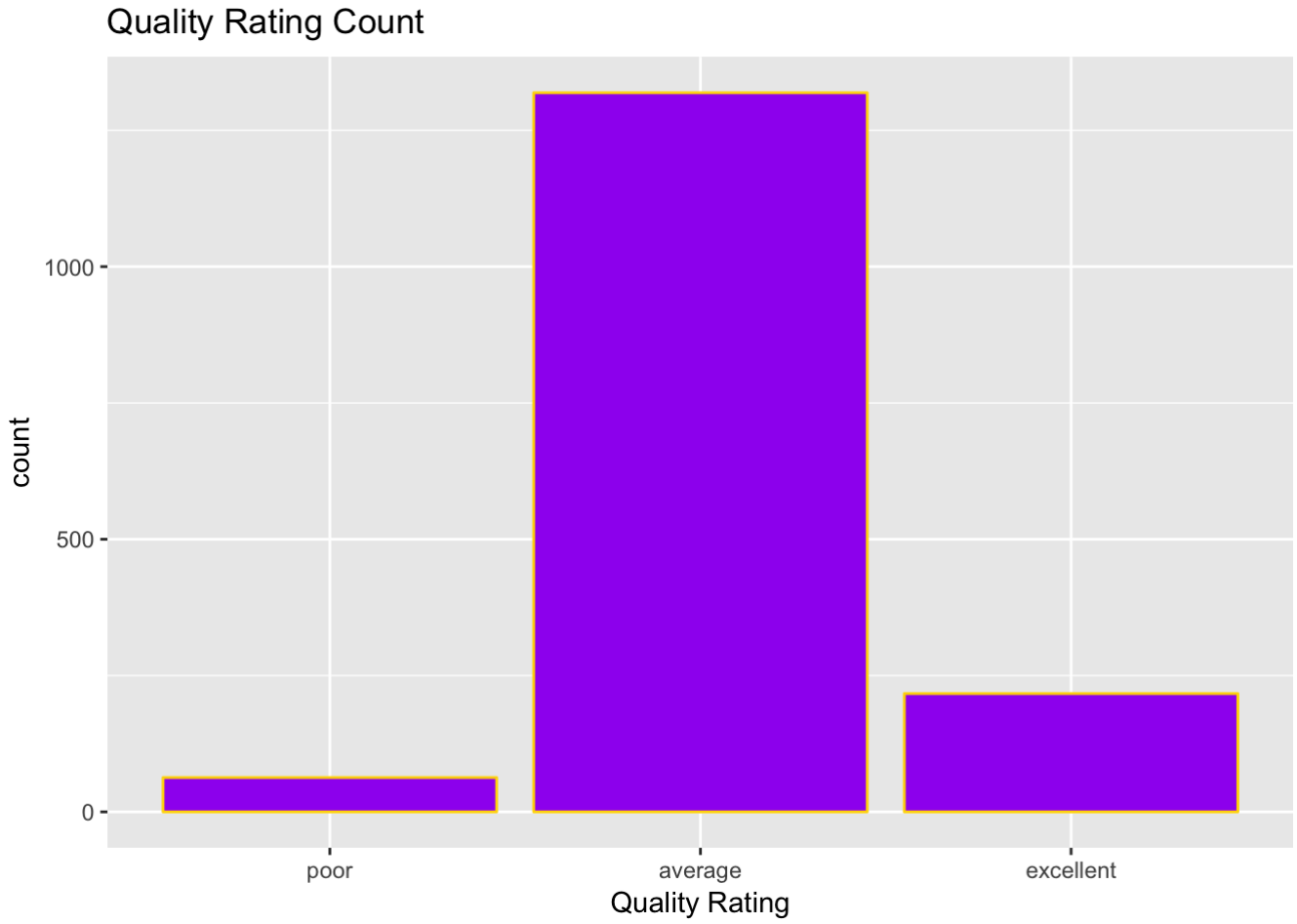
**

Were there any interesting or surprising interactions between features?

The plot involving sulphates and volatile acidity is pretty interesting. It was an impressive graph because it clearly showed a trend between the subqualities that couldn't go unnoticed. the better the subquality, the lower the points on the graph went. in this specific graph, excelent wines had lower sulphates and lower volatile acidity than it's othe subquality counterparts. The relationship between citric acid and alcohol is considered important as well.

Final Plots and Summary

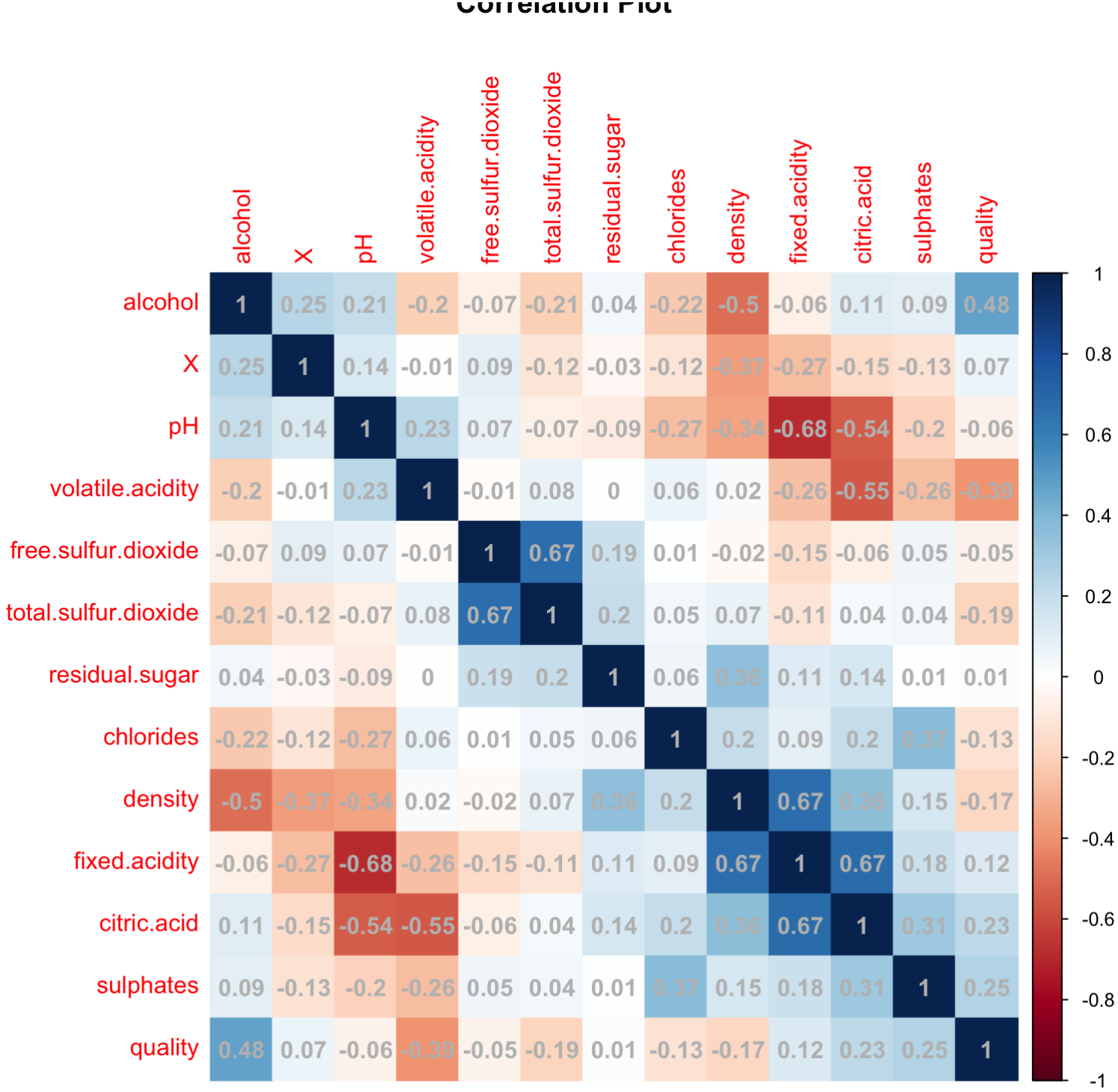
Plot One



Description One

Plot One was intersting because it basically shows how wine judgers are quick to give the quality of red wine a score of either 5 or 6, which would be considered average. Based on this data set, it shows that there were very few poor wines and excellent wines. A wine quality score of 7 or greater is considered excellent and a great achievement.

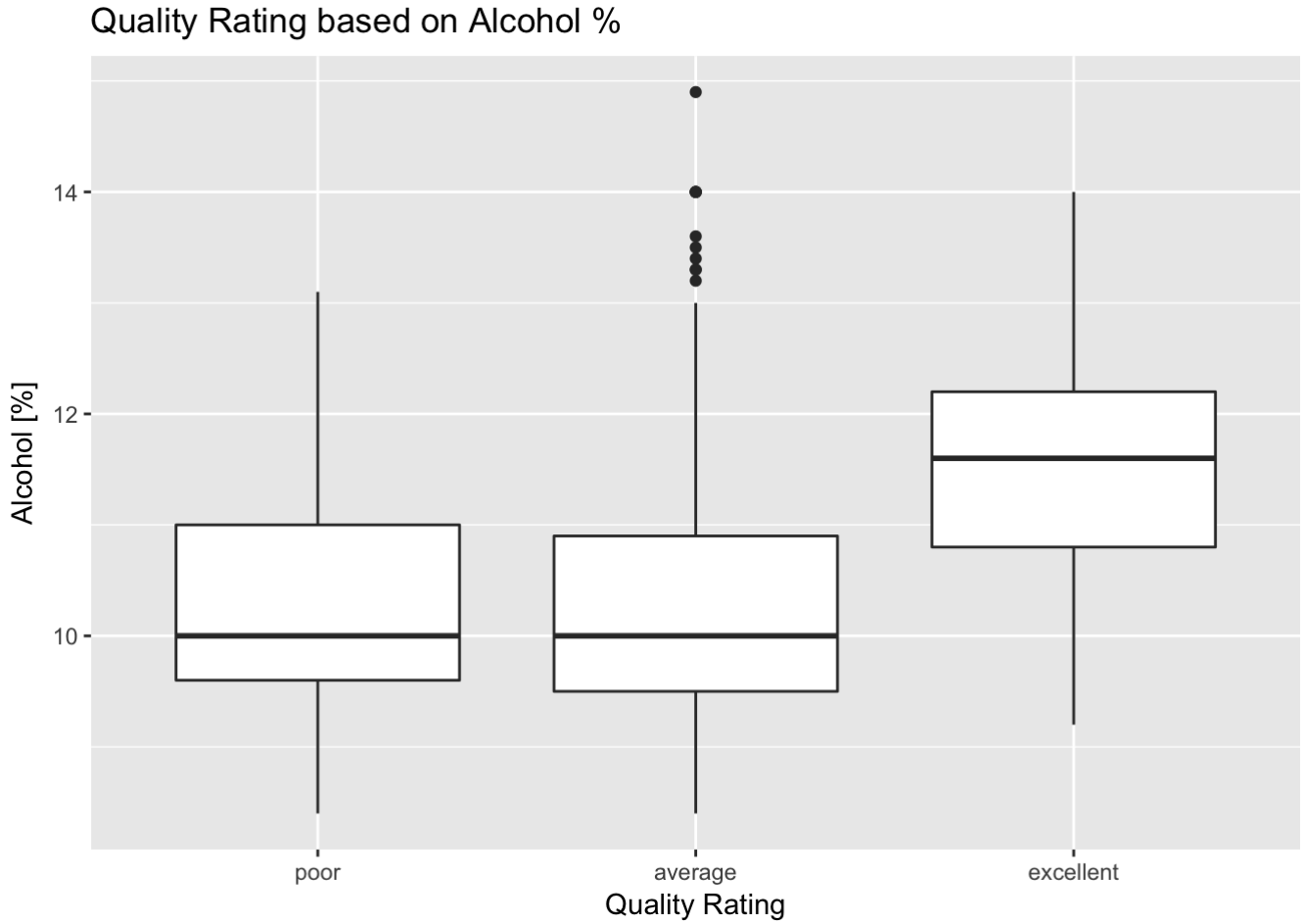
Plot Two


Correlation Plot
alcohol
X
pH
volatile.acidity
free.sulfur.dioxide
total.sulfur.dioxide
residual.sugar
chlorides
density
fixed.acidity
citric.acid
sulphates
quality
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Description Two

Plot 2 describes a beautiful correlation plot between all variables, and is why I chose it. My main interest was quality and it was interesting to see that it had considerable correlations with the variables volatile acidity, alcohol, citric acid, and sulphates. There were also positive correlations between variables that weren't initially my main focus. Those correlations were: free sulfur dioxide and total sulfur dioxide, fixed acidity and density, and fixed acidity and pH.

Plot Three



Description Three

Plot three is a boxplot of alcohol based on quality rating and it basically shows how alcohal content played a big role in the rating. excellent quality ratings, which are scores 7 and above had a mean of 11.52 compared to the means of poor and average scores of 10.22 and 10.25 respectively.

Reflection

The red wine data set contained 1,599 observations of red wine along with 13 variables. Throughout this project, I wanted to discover what drove the quality of wine and what factors took place that had much to do with the wine's quality score. I've learned that most wine judgers gave the red wine in the dataset scores of either 5 or 6, which is consdered average.

Alcohol content appears to be the biggest factor in determing excelent wine, which was the most surprising in my opinion. It's safe to say that wine judgers like larger amounts of alcohol content in their wine, which causes more euphoria.The amount of citric acid in wine also make a big difference for quality. The more citric acid the wine has the more likely it would also have a considerable amount of alcohol content.

Volatile Acidity in wine plays a role in quality as well. Poor wines tend to have high volatile acidity scores and becomes lower for each subquality. This means that too much volatile acidity is bad for wine, period. Average wines have a lower volatile acidity than poor wines, but have higher rates of sulphates. Out of all subquality ratings, excellent red wines had lower sulphates and low volatile acidity. This would be a revelation for determining what really makes a poor red wine and great wine.

Believe or not, criteria for anything is usually biased because it was created by the predferences of people. The criteria for this dataset and what caused excellent wine scores could be totally different from other datasets when making discoveries about the data. Having a bigger sample of red wine data could possibly shed some light on what truly plays a role in great red wine. Surveys from customer populations would beinteresting because it gives a different perspective on the matter. not only does it give a different perspective, but it can also predict what customers prefer when purchasing red wine.

Any Struggles?

There wern't as many strong correlations between a wide variety of variable in this dataset. In my opinion, I believe that there could/should be more variables that could support the factors that support qualty for red wine. Luckily, it wasn't too difficult to see what contributes to red wine because of the correlation plot.